Digital Signal Processing Constrained temporal structure for text-dependent speaker verification

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

Constrained Temporal Structure for Text-Dependent Speaker Verification

Anthony Larcher[a,1,*], Jean-Francois Bonastre[a], John S.D. Mason[b]

[a] *University of Avignon LIA-CERI, 84911 Avignon Cedex 9, France*
[b] *Speech and Image Research, School of Engineering, Swansea University, Swansea SA2 8PP, U.K.*

Abstract

In the context of mobile devices, speaker recognition engines may suffer from ergonomic constraints and limited amount of computing resources. Even if they prove their efficiency in classical contexts, GMM/UBM systems show their limitations when restricting the quantity of speech data. In contrast, the proposed GMM/UBM extension addresses situations characterised by limited enrolment data and only the computing power typically found on modern mobile devices. A key contribution comes from the harnessing of the temporal structure of speech using client-customised pass-phrases and new Markov model structures. Additional temporal information is then used to enhance discrimination with Viterbi decoding, increasing the gap between client and imposter scores. Experiments on the MyIdea database are presented with a standard GMM/UBM configuration acting as a benchmark. When imposters do not know the client pass-phrase, a relative gain of up to 65% in terms of EER is achieved over the GMM/UBM baseline configuration. The results clearly highlight the potential of this new approach, with a good balance between complexity and recognition accuracy.

Keywords: Speaker recognition, Text-dependent, Password, Embedded application

[*] Corresponding author
Email addresses: anthony.larcher@univ-avignon.fr (Anthony Larcher), jean-francois.bonastre@univ-avignon.fr (Jean-Francois Bonastre), j.s.d.mason@swan.ac.uk (John S.D. Mason)
[1] Present address: alarcher@i2r.a-star.edu.sg

1. Introduction

The speech signal offers several advantages over other biometric signals, with distinct benefits coming from the potential to link together information derived from the context and the content of the message as well as the voice biometric itself. With appropriate classification and fusion, these components can be brought together to enhance any biometric validation process.

There are however practical constraints within the cell-phone scenario, stemming largely from ergonomic factors and the available computing resources typically found within such hand-held devices. Such embedded applications impose constraints and an important one in terms of recognition performance can be the quantity of data, particularly for reference models but also for the subsequent test phase. In certain applications these quantities might prove to have a critical influence on recognition accuracy. Examples include security systems that might well have speech of only a few words spanning just 2 or 3 seconds. The main contributions of this paper address these issues with new computational structures designed to harness maximum information from the temporal structure information (TSI) of speech to reinforce the acoustic modelling.

Classical speaker recognition engines offer a high level of performance as shown for example during NIST evaluations [1]. However such systems, which are invariably founded on the GMM/UBM paradigm [2], exhibit high sensitivity to the quantity of data, particularly the reference model data [3], [4], [5]. Their performance degrades strongly while reducing the duration of speech material available [6, 7, 8]. For situations where the speech duration is below 30 seconds, recognition performance falls rapidly [9, 10]. Text-dependency is well known to compensate for the lack of data by constraining the acoustic content of the spoken utterance [11].

Meaningful comparison of recognition accuracy in text-dependent speaker verification tends to be very difficult due to the lack of controlled evaluations and large scale databases, essential particularly when error rates are very low [12]. Hence the tendency of the community towards the text-independent scenario that benefits from the NIST large scale databases and the independent evaluations [1]. However, two major trends dominate the field of text-dependent speaker verification. Approaches based on dynamic programming have been proposed for tasks where the quantity of speech is limited [13, 14, 15, 16]. They provide a precise modelling of the time constraints but lack the generalization power available with hidden Markov model (HMM) approaches [17]. Indeed, HMM and GMM models which are the most common modelling methods [18] are more robust to speaker or environment variabilities and can take advantage of larger amounts of data [11]. Depending on the type of application that is targeted, HMMs can be used to model whole sentences [19, 20], word-level units [21, 22] or phone-level units [23, 24]. In addition to the two major approaches that dominate text-dependent speaker verification, text-dependent and text-independent speaker verification often cross-pollinate each other. While there have been several attempts to adapt Support Vector Machines [25, 26] or i-vector systems [27, 28] to take advantages of a lexical constraints, others incorporate text-dependent techniques in text-independent applications such as [29, 30].

In this paper, a classification structure that takes advantage of the temporal structure of the speech utterance is examined. The structure utilises text-dependencies derived from a multilayer classifier illustrated in Figure 1, the foundation of which is the standard GMM/UBM. These first two layers are complemented by a third layer that harnesses temporal structure information extracted from speaker-specific phrases. The approach, described in [31], [32], [33] and further developed in this paper, takes advantage of the temporal structure of pass-phrases, an example of which is "Ce petit canard apprend à nager[2]". In order to model the TSI of such a pass-phrase while achieving statistical modelling from the GMM/UBM, we propose to extend the standard paradigm with an HMM/Viterbi approach. Finite-state models aim to incorporate pass-phrase-based information, like temporal organisation of acoustic features, not otherwise harnessed by classical GMM/UBM approaches.

A key point here is the inclusion of additional temporal information within the finite-state modelling. This additional information is used to constrain the Viterbi decoding in order to enhance discrimination. It does so by using the temporal structure of the given pass-phrase. A set of N classical HMM nodes is arranged in time sequence with the transitions in time from one node to the next controlled first by the normal acoustic features and then by additional temporal information.

The proposed structure is designed specifically to accommodate the use of two such simultaneous synchronous signals. The roles of the two can be clearly separated: first, variants of the conventional GMM/UBM nodes; and

[2] This little duck is learning to swim.

second, additional synchronization control of state transitions. Here, the latter comes from the acoustic signal and divides each pass-phrase into segments overarching several states of the HMM, as shown in Figure 2. These overarching segments provide constraints at the lexical level that are in addition to those of the finite-state models and that can be harnessed by the recogniser. We refer to these as lexical constraints.

The approach proposed in this work is related to that of others in the literature. For example in [34] Becerra Yoma and Facco Pegoraro constrain the state duration of word-units HMMs. Additional knowledge is included in the HMM topology by training different transition probabilities depending on the position of a given word in the speech segment. In [35], speaker dependent semi-continuous HMMs are compared to a reference HMM to produce a discriminative representation of the speaker pronouncing a given pass-phrase. In the two previous works, the use of a background HMM to adapt the speaker dependent models or to model the alternative hypothesis strongly limits the flexibility of the system in terms of lexicon. The architecture proposed in this work takes advantages of the GMM/UBM framework to model the alternative hypothesis and adapt the speaker model and thus, gives more flexibility to the user to choose a specific pass-phrase. Another related work is proposed in [36] where supra-segmental temporal information is used to reinforce the robustness of a Dynamic Time Warping algorithm. By combining the different information sources in a later stage, using a neural network, the system does not take advantage of the temporal synchronization of the different signals as is the case in the work presented here.

Many techniques exist in the literature to compensate for the variabilities due to channel or environment mismatch in the GMM/UBM framework. Some of these techniques like RASTA and Short Term Gaussianization work at the parametrization level [37, 38] when others are dedicated to score normalization [39, 40]. These techniques have not been applied in this article which focus on the advantages of our approach compared to the GMM/UBM. Nevertheless, most of the techniques that have been developed for the GMM/UBM may be applied to our approach and are expected to provide similar improvement.

The overall system architecture is described in Section 2. The impact of the lexical information in constraining the Viterbi decoding is described in Section 3. Section 4 describes the experimental protocol and results. It includes a description of the MyIdea database [41]. Section 5 summarises the benefits of this approach and presents future work directions.

2. A three-level acoustic architecture

The architecture presented in Figure 1 is an extension of the standard GMM/UBM paradigm. Throughout the text we refer to this new structure as the Embedded LIA_SpkDet[3] (EBD) [42]. This architecture is configured to deal with a user-customised speaker recognition task. Each client has a customised pass-phrase, which is unique to that person. Hence some form of text dependency can be harnessed within the speaker recognition system.

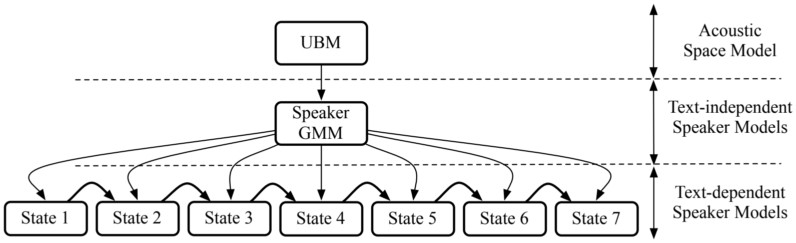


Figure 1: General view of the *EBD* architecture.

2.1. Training phase

The two first layers of the EBD consist of a classical GMM/UBM speaker recognition system. The upper layer, a standard universal background model (UBM), aims to model the acoustic speech space. This GMM is built off-line using a suitably large amount of data and the classical EM/ML algorithm [2]. A text-independent speaker-specific GMM (2^{nd} layer) is then adapted from the UBM for each client speaker with the client data using the EM algorithm and the maximum a posteriori (MAP) criterion [43].

Finally, a semi-continuous hidden Markov model (SCHMM) [44] is used with the goal of harnessing the TSI of the individual pass-phrase. This third layer introduces text-dependency into the client models. In order to initialize the S states of the SCHMM, the pass-phrase is cut into S segments $\{seg_i\}$ of equal length. The training algorithm is then a two step process. During the first step, a GMM is trained for each segment seg_i by adapting only the weight parameters of the text-independent client model. For the c^{th}

[3] Implementation of the EBD is based on the open-source toolkit LIA_SpkDet, part of ALIZE, `http://alize.univ-avignon.fr`

distribution of the GMM, the adapted weight parameter $\hat{w}_c$ is given by:

$$\hat{w}_c = \left[\alpha_c \frac{n_c}{T} + (1 - \alpha_c) w_c\right] \gamma \tag{1}$$

where w_c is the weight of the c^{th} distribution of the text-independent client model, n_c is the occupancy of distribution c, T is the total number of speech frames allocated to the c^{th} distribution and γ assures that the weight parameters of a GMM state sum to one. Finally, α_c is given by

$$\alpha_c = \frac{n_c}{n_c + r} \tag{2}$$

with r a regulation factor empirically determined. The second step consists of running a Viterbi decoding with the current SCHMM on the enrolment data to produce new segments $\{seg_i\}$. The two steps are repeated until convergence of the segmentation. During this process, the number of states is systematically reduced from a given maximum by removing those states that receive fewer than a pre-set minimum number of frames. The optimal initial number of states and the minimum number of frames per state are experimentally determined. Such a combined system was originally proposed in [17] for speaker recognition and extended to word recognition in [45].

2.2. Testing phase

During a test, two scores are computed. The first is the conventional GMM/UBM log-likelihood ratio obtained from the text-independent speaker model; the second is from the SCHMM. For the case of the original GMM/UBM, the log-likelihood of a test sequence $\mathcal{O} = \{o_t\}, t \in [1, T]$ and a N-distribution GMM is given by:

$$\mathcal{L}(\text{GMM}) = \frac{1}{T} \sum_{t=1}^{T} \log \left(\sum_{c=1}^{N} \gamma^c \mathcal{N}(o_t | \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c) \right) \tag{3}$$

where γ^c, $\boldsymbol{\mu}^c$ and $\boldsymbol{\Sigma}^c$ are respectively the weight, mean vector and covariance matrix of the c^{th} distribution of the GMM and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Similarly, for the case of the SCHMM, the log-likelihood of the test sequence $\mathcal{O}$ on the SCHMM is:

$$\mathcal{L}(\text{SCHMM}) = \frac{1}{T} \sum_{t=1}^{T} \log \left(\sum_{c=1}^{N} \gamma_t^c \mathcal{N}(o_t | \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c) \right) \tag{4}$$

where γ_t^c is the weight parameter of the c^{th} distribution of the SCHMM state allocated by the Viterbi decoding at time t. Note that the only difference between the two log-likelihood expressions is the parameter γ that weights the contribution of each distribution. Indeed, all distributions from the text-independent speaker model are shared among model states but their weights are unique for each state. Log-likelihoods computed on the GMM and SCHMM (Eq. 3 and 4) are normalized by using the UBM to form log-likelihood ratios that are linearly combined. The final score for the decision stage is given by:

$$\mathcal{S} = \Bigg[\lambda \times \mathcal{L}(\text{SCHMM}) + (1-\lambda) \times \mathcal{L}(\text{GMM})\Bigg] - \mathcal{L}(\text{UBM}) \qquad (5)$$

where $\mathcal{L}$(UBM) is the log-likelihood of the test sequence $\mathcal{O}$ on the UBM and λ is a empirically chosen in $[0; 1]$.

In addition to the TSI introduced by the semi-continuous HMM, further information can be added to the system in the form of overarching temporal constraints as illustrated in the lower part of Figure 2 (three segments in this case). In Section 3 we describe such additional constraints.

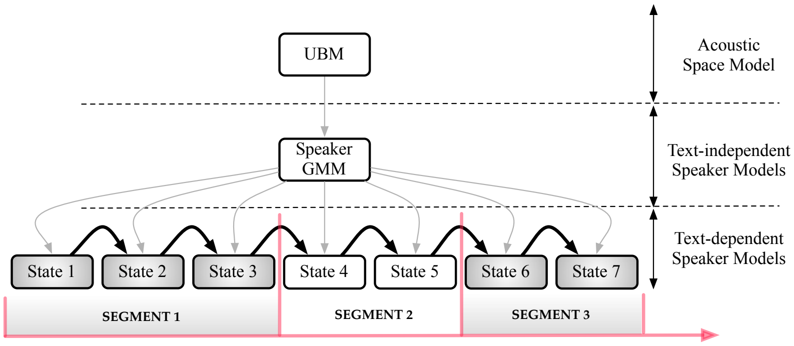


Figure 2: General view of the *EBD* architecture integrating additional TSI in the form of three overarching segments.

2.3. Complexity

With the given mobile application in mind it is important to consider the complexity of the proposed approach and in particular that of the testing phase; the training is less important since that can be performed offline. We estimate the complexity of the testing phase and make comparison with both a standard GMM/UBM and an equivalent HMM.

The complexity is very much dominated by two tasks, namely the likelihood computation and the Viterbi decoding. The likelihood computation is likely to be dominated by the exponential function, for which an estimation is $O(n^{\frac{1}{2}}M(n))$ where n is the number of digits of precision and $M(n)$ is the complexity of the multiplication algorithm. Other key parameters are the number of acoustic features, T, the dimension D of those features and the number of Gaussian components, C. Thus a reasonable estimate of the computational load would be $O(n^{\frac{1}{2}}M(n)CDT)$. Complexity of the Viterbi algorithm can be expressed as $O(TS^2)$ where S is the number of states in the model.

The computational effort associated with an HMM is dominated largely by two components, namely the likelihood computation and the Viterbi decoding. In our EBD architecture, the likelihood component is reduced to that of the GMM/UBM structure since the states of the SCHMM share all components with the text-independent speaker model of the second layer. Thus additional complexity of our approach compared to that of the GMM/UBM is restricted to the Viterbi decoding only.

In assessing the system complexity, we have examined the likelihood/Viterbi ratio with two approaches, namely complexity estimates and profiling of the program. The first is a theoretical estimate of complexity that leads to likelihood/Viterbi ratio of $32 \times n^{\frac{1}{2}}M(n)$. The second relates to the actual computation load and leads to an estimate of a likelihood/Viterbi ratio of 16:1. From the combination of these two observations it could reasonably be concluded that likelihood computation itself dominates overall and thus the additional cost of our approach over the standard GMM/UBM is negligible.

3. The constrained Viterbi

We reinforce the TSI modelling of the client-specific utterance by using additional temporal information. The bottom layer SCHMM of the EBD architecture is based on a conventional left-to-right topology for which two types of transitions are defined (Figure 3): transitions $\boldsymbol{E}$ that are fixed and set to equiprobability and transitions $\boldsymbol{S}$, the values of which vary during the decoding. Each transition $\boldsymbol{S}$, with a default probability of 0, is driven by a discrete event or synchronization point. During the decoding, the value of a transition $\boldsymbol{S}$ turns to 1 when reaching the associated synchronization point and back to 0 once this point is past. This temporal constraint forbids the Viterbi path to go across certain areas of the lattice (areas illustrated on Figures 4(a) and 4(b) by filled states). In this context, the bottom layer of the EBD architecture could be compared to a succession of sub-SCHMMs and the Viterbi path is then computed from one synchronization point to the next with the corresponding sub-SCHMM. The synchronization constraint which is applied in the training and test phase could be linked to an external source, leading to the notion of synchronization between two sources of information.

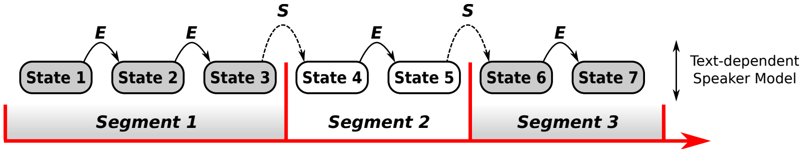
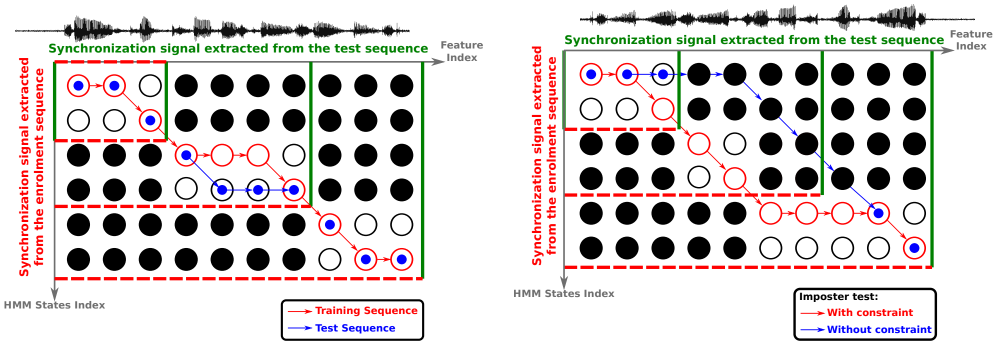


Figure 3: Use of a lexical synchronization in the bottom layer of the EBD architecture constrains the Viterbi decoding and increases the discriminative power of the system. In this example, the synchronization signal consists of 2 discrete points separating 3 overarching segments. Transitions $\boldsymbol{E}$ are fixed and set to equiprobability while transitions $\boldsymbol{S}$ value varies during the decoding.

The addition of complementary information coming from the synchronization points could improve the training of SCHMM models. Such an improvement could potentially lead to increased 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.

The effect of the lexical constraint is illustrated in Figures 4(a) and 4(b). 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 segmentations computed on both the training and the testing utterances.

Figure 4(a) shows paths obtained on a same model for two different pass-phrase occurrences: the occurrence used to train this SCHMM model and one occurrence of the same pass-phrase, pronounced by the client. The temporal structure of a client test utterance is assumed 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 allowed zones only. This is


Synchronization signal extracted from the test sequence
Feature Index
Synchronization signal extracted from the enrolment sequence
HMM States Index
Training Sequence
Test Sequence
Synchronization signal extracted from the test sequence
Feature Index
Synchronization signal extracted from the enrolment sequence
HMM States Index
Imposter test:
With constraint
Without constraint

(a) Viterbi alignment of client training and test utterances with a lexical constraint applied during the decoding.

(b) Alignment of imposter utterance with and without lexical constraint.

Figure 4: Schematic drawing of the alignment of acoustic feature sequences for different configurations. Time dimension is represented on the horizontal axis while progress across the HMM model is given by the vertical axis.

illustrated in Figure 4(a). In this case, the synchronization constraint has no effect on the Viterbi decoding and the resulting score remains high.

Figure 4(b) shows two expected alignment paths for a given imposter test utterance, one with the synchronization constraint (the more central one) and one without the synchronization constraint. The temporal structure of an imposter utterance is assumed to be different from that of the true client pass-phrase structure. The path resulting from the alignment of this utterance on the client pass-phrase SCHMM without applying the lexical constraint is assumed to go through the forbidden area. Thus, the synchronization constraint forces the algorithm to find a path through the allowed zones. This path is not optimal and the imposter score is therefore lower than the one computed without any constraint.

4. Experiments

This section presents the experiments designed to evaluate the benefits of both the SCHMM extension and the constrained decoding. It includes a description of the evaluation corpus and the experimental protocol.

4.1. The MyIdea database

The multi-modal MyIdea database [41] includes recordings of 30 male speakers, each pronouncing 25 sentences in 3 sessions under controlled acoustic conditions. Twelve of the sentences, common to all speakers and sessions, are used for our experiments[4]. Out of these twelve sentences, ten are approximately three seconds long and will be referred to as 3s sentences, two are 6s long. The ten 3s sentences are used as user-customized pass-phrases.

4.2. Experimental protocol

The 30 males of the MyIdea database are separated into two groups of 15 speakers each. The whole recorded material of each group is successively used to train the UBM using the classical EM algorithm [46] while the 15 speakers of the remaining group are used for enrolment and tests. Note that when a group of speakers is dedicated to UBM training, data from these speaker is not used for any other purpose. Due to the lack of data, a jackknifing process is used within the enrolment and test group by successively considering each speaker as a client while each of the 14 other speakers of the group is regarded as an imposters [31].

Each client GMM (layer two of the EBD architecture) is derived using two 6s sentences and one occurrence of the selected 3s sentence (around 15 seconds of speech in total). The given 3s sentence is also used to train the client pass-phrase SCHMM. Over the two groups of 15 speakers, 900 SCHMM models are trained (Table 1).

A speaker model is trained using one 3s sentence from one of the 3 sessions. The same 3s sentence from the two other sessions of the same speaker is used for target trials, giving 2 target trials per speaker model. This process, applied to the 900 target models obtained after jackknifing, leads to a total of 1,800 target trials over the two groups of speakers.

For each speaker model, imposter tests are computed against the 3s sentences from the others 14 speakers of the same group. Three configurations of *imposter* tests are considered.

UNKNOWN configuration the linguistic content of the imposter test occurrences is *different* from that of the material used to train the models. Each client model is compared to *three randomly selected* 3s sentences

[4] The others 13 sentences differ across speakers and sessions and can thus not be used for text-dependent experiments.

Table 1: Number of GMM speaker models trained per sentence, session, speaker and group for the MyIdea database. Note that the number of SCHMM models (layer three of the EBD architecture) is equal to the number of GMM speaker model given in this table.

	Number of models	Detail	
per 3s sentence	1	1	
per speaker per session	10	1	×10 sentences
per speaker	30	(1×10)	×3 sessions
per group	450	$(1 \times 10 \times 3)$	×15 speakers
Total	900	$(1 \times 10 \times 3 \times 15)$	×2 groups

(one per session) *out of the 9 remaining sentences* of each of the 14 imposter speakers. This reflects the condition when an *imposter does not know* the client pass-phrase.

KNOWN configuration the linguistic content of the imposter test occurrences is *the same* as that of the material used to train the models. Each client model is compared to the *three 3s sentences* (one per session) of each of the 14 imposter speakers. This reflects the condition when an *imposter knows* the client pass-phrase.

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

For both KNOWN and UNKNOWN configurations, the number of imposter tests is constant across the clients. Moreover, the global number of imposter tests is 37,800 in the two first configurations and 75,600 in the ALL configuration. Further experimental details are presented in [31].

4.3. System configuration

Each file of the database is parametrized into a sequence of 32-dimensional vectors made up of 15 Linear-Frequency Cepstral Coefficients (LFCC), the log-energy and the corresponding Δ coefficients. An energy labelling is applied to separate speech frames from the non-speech frames and cepstral mean and variance normalization (CMVN) is applied to the remaining frames [47].

The number of components in GMMs is fixed to 256 across the EBD architecture. The initial number of states in SCHMM to model the 3s sentences is set to 20 regardless of the sentence and the regulation factor r of the MAP adaptation is set to 14. At the end of the iterative training, over all 3s sentences of the 30 speakers, the average number of states per sentence is just over 19. For comparison, modelling the same ten 3s sentences of the MyIdea database by using non-contextual phone models requires an average of 24.2 states per sentence.

4.4. Results

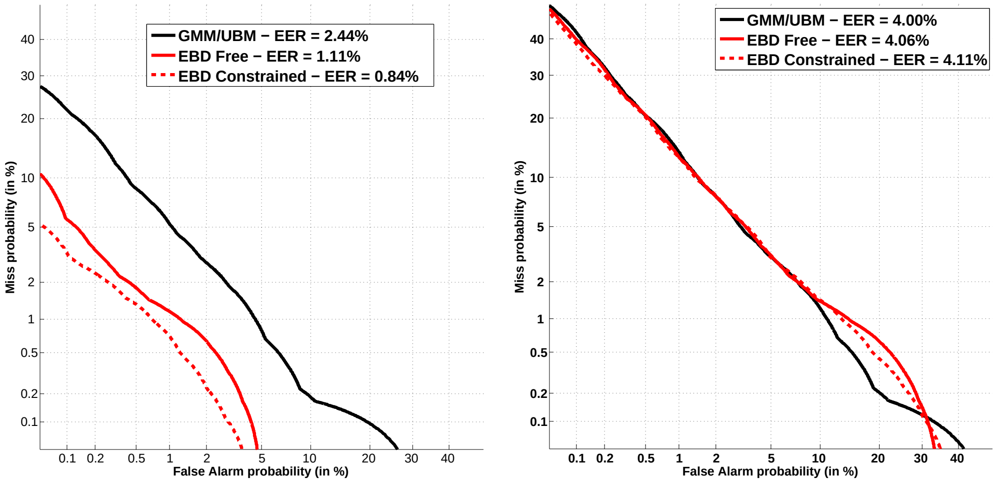
Experiments aim to assess the contributions of the three components, namely the GMM/UBM, the SCHMM and the constrained Viterbi. The GMM/UBM is regarded as the baseline. The experimental results are presented in Table 2.

Table 2: EER (%) of a standard GMM/UBM configuration compared to the EBD system with 20 states per SCHMM, with the free or constrained Viterbi alignment.

Configuration	GMM/UBM baseline	EBD system % EER	
	% EER	Free	Constrained
UNKNOWN	2.44	1.11	0.84
KNOWN	4.00	4.06	4.11
ALL	3.22	2.83	2.89

The main advantage expected from the EBD system compared to a classical GMM/UBM is the incorporation of pass-phrase-based information as well as the pass-phrase itself and the relative TSI. This hypothesis is supported by results in Table 2 which shows error rates fall from 2.44% to 1.11% using the SCHMM when imposters do not know the client pass-phrases. When the imposters know the client pass-phrases (KNOWN) the performances of the EBD and GMM/UBM systems are equivalent.

Additionally, Figures 5(a) and 5(b) show that this hypothesis stands for different miss probability and false alarm ratios as the DET curves of the EBD system are well below those of the GMM/UBM when imposters do not know the client pass-phrases but comparable when imposters know the client pass-phrase.



(a) UNKNOWN - condition

(b) KNOWN - condition

Figure 5: Detection Error Trade-off curves of the GMM-UBM baseline and EBD system with 20 states per SCHMM, with free or constrained Viterbi alignment. Cases where the imposters do not know the client pass-phrase and pronounce a different utterance (UNKNOWN - condition) and where the imposters know the client pass-phrases (KNOWN - condition)

In order to evaluate the effect of the lexical constraint, a new experiment is performed using the constrained Viterbi decoding. Here, knowledge of the correct client-specific pass-phrase is used in constraining the alignment. The lexical constraint consists of five synchronization points corresponding to word boundaries extracted from each enrolment and test utterance using the LIA SPEERAL Toolkit [48]. The fixed number of synchronization points and their selection process has been empirically determined and various strategies could be applied in the future to optimize this process. The results of this experiment are presented in the fourth column of Table 2. As expected, the performance of the EBD in the UNKNOWN condition improves when constraining the Viterbi alignment with the lexical synchronization. Indeed, the EER drops by 65% relatively to the baseline GMM/UBM (from 2.44% to 0.84%). Here, imposters do not know the client pass-phrase and pronounce a different utterance whose temporal structure is penalised by the synchronization constraint.

Figure 6-A shows that the client score distributions with and without

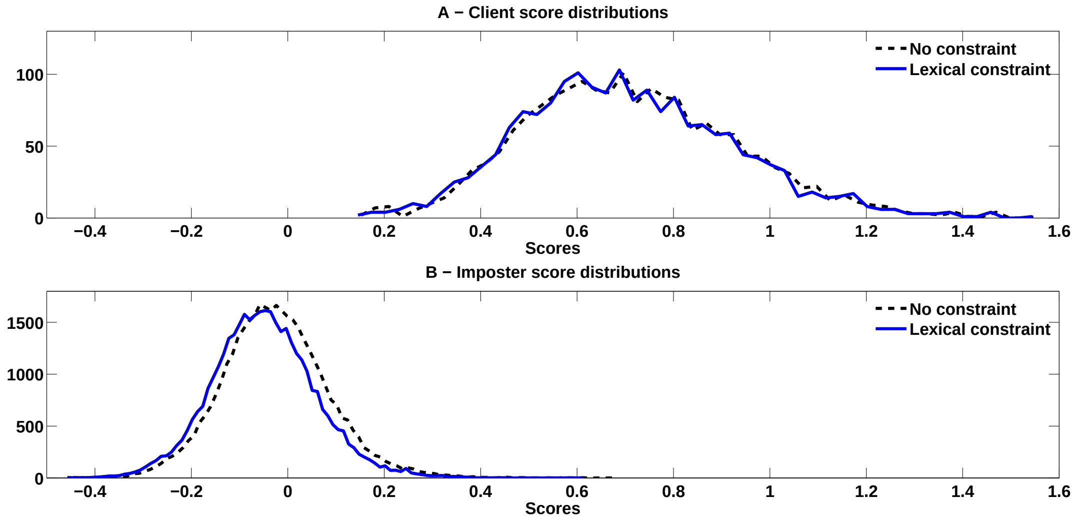


Figure 6: Evolution of the client and imposter text-dependent score distributions when using additional information coming from a lexical synchronization (constrained Viterbi decoding) or no synchronization information (classical Viterbi decoding). Case where imposters do not know the client pass-phrase and pronounce a different utterance (UNKNOWN).

constraining the Viterbi decoding are essentially unchanged. This might be explained by any potential information coming from a lexical synchronization being highly correlated with the acoustic information already exploited by the SCHMM extension. At the same time, these results show that the constrained Viterbi is effective in that it does not degrade client scores.

However, benefits are seen in the case of imposters. Figure 6-B confirms that imposter scores decrease consistently, in terms of the distributions. Furthermore, Figure 7 shows that the Viterbi constraint affects a very large number (95%) of imposter tests. This shows that the synchronization constraint leads to sub-optimal alignment when the temporal structure of the test utterance is different from the training pass-phrase structure.

5. Conclusions and future work

The approach proposed in this paper addresses the issue of speaker recognition in the challenging context of having only a limited quantity of data

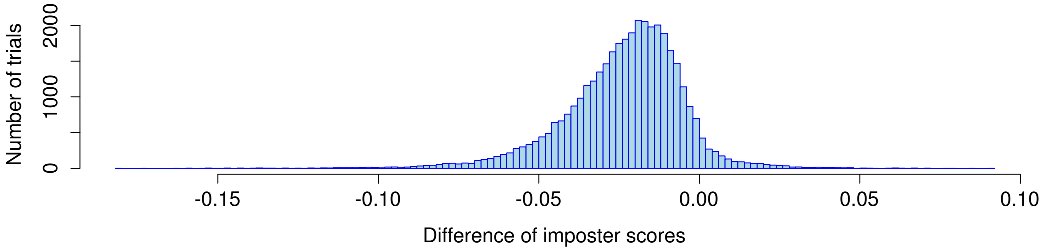


Figure 7: Distribution of imposter score differences, between Viterbi decoding with and without lexical constraint, when imposters do not know the client pass-phrase and pronounce a different utterance (UNKNOWN).

during both training and test phases. A new temporal structure combining the GMM/UBM text-independent paradigm and the HMM/Viterbi abilities has been investigated, aimed at harnessing the temporal structure information of user-customised pass-phrases. This new three level architecture, referred to as the EBD, is structured so that constraints can be applied to state transitions.

Experimental results are presented for the baseline GMM/UBM and for the EBD, with and without Viterbi constraints. The performances for the EBD and the EBD constrained show relative improvement of 65%, with EERs from the baseline of 2.44% to 1.11% to 0.84% when imposters utter a passphase other than the correct and client-specific phrase. These results show the benefits of harnessing the temporal structure information in combination with the classical GMM/UBM configuration.

Even if the EBD demonstrates its ability to take advantage of the temporal structure of the speaker pass-phrase, the results remains very similar to those of the GMM/UBM when imposters know the client pass-phrase (see row marked 'KNOWN' in Table 2). In [34], Becerra Yoma and Facco Pegoraro conclude that state duration restriction seems not to be relevant in the context of clean speech when imposters know the client pass-phrase but that it can lead to significant improvement in noisy conditions. This conclusion may explain our observation and further experiments have to be conducted in the future on noisy signal to evaluate the impact of a temporal structure for the case where imposters know the client pass-phrase. A deeper analysis of the client-specific temporal information may also help to further optimize the EBD for the KNOWN condition, for example by taking into account intrinsically the state duration information as done in [34].

Another source of potential improvement could come from the substitution of the lexical synchronization by a less correlated source of information. Future work will focus on multi-modality by replacing the lexical synchronization with temporal information extracted from the video part of the audio-visual stream to increase the performance of the current approach. Additionally, the use of a second modality into the verification process could be very useful to thwart synthetic playback attack by monitoring the temporal correlation of the two streams of information.

Moreover, given that the first results show the ability of the EBD to take advantage of the linguistic content of speaker-specific pass-phrases, more tests are to be performed to evaluate the performance of the EBD system with more utterance-variability, for example considering the pass-phrase duration.

References

[1] A. F. Martin, C. S. Greenberg, NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels, in: Annual Conference of the International Speech Communication Association (Interspeech), 2009, pp. 2579–2582.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, D. A. Reynolds, A tutorial on text-independent speaker verification, EURASIP Journal on Applied Signal Processing 4 (2004) 430–451.

[3] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre, J. S. Mason, Influence of task duration in text-independent speaker verification, in: Annual Conference of the International Speech Communication Association (Interspeech), 2007, pp. 794–797.

[4] B. Fauve, N. Evans, J. S. Mason, Improving the performance of text-independent short duration SVM-and GMM-based speaker verification, in: Speaker and Language Recognition Workshop (Odyssey), 2008, pp. 1–7.

[5] R. Vogt, S. Sridharan, Minimising speaker verification utterance length through confidence based early verification decisions, in: Proceedings of

the Third International Conference on Advances in Biometrics, Springer, 2009, pp. 463–472.

[6] A. Kanagasundaram, R. J. Vogt, D. Dean, S. Sridharan, PLDA based Speaker Recognition on Short Utterances, in: Speaker and Language Recognition Workshop (Odyssey), 2012, pp. 1–6.

[7] T. Stadelmann, B. Freisleben, Dimension-Decoupled Gaussian Mixture Model for Short Utterance Speaker Recognition, in: International Conference on Pattern Recognition (ICPR), 2010, pp. 1602–1605.

[8] M. Nosratighods, E. Ambikairajah, J. Epps, M. J. Carey, A segment selection technique for speaker verification, Speech Communication 52 (9) (2010) 753–761.

[9] B. Fauve, Tackling variabilities in speaker verification with a focus on short durations, Ph.D. thesis, School of Engineering Swansea University (2009).

[10] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, M. Mason, I-vector Based Speaker Recognition on Short Utterances, in: Annual Conference of the International Speech Communication Association (Interspeech), 2011, pp. 2341–2344.

[11] M. Hébert, Text-dependent speaker recognition, Springer-Verlag, Heidelberg, 2008.

[12] G. R. Doddington, Speaker recognition evaluation methodology–an overview and perspective–, in: Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), 1998, pp. 20–23.

[13] S. Furui, Cepstral analysis technique for automatic speaker verification, IEEE Transactions on Acoustics, Speech, and Signal Processing 29 (2) (1981) 254–272.

[14] R. P. Starpert, J. S. Mason, A segmental mixture model for speaker recognition, in: European Conference on Speech Communication and Technology (Eurospeech), Vol. 4, 2001, pp. 2509–2512.

[15] V. Ramasubramanian, A. Das, V. Kumar, Text-Dependent Speaker-Recognition Using One-Pass Dynamic Programming Algorithm, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2006, pp. 901–904.

[16] B. Avinash, S. Guruprasad, B.Ygnannarayana, Exploring subsegmental and suprasegmental features for a text-dependent speaker verification in distant speech signals, in: Annual Conference of the International Speech Communication Association (Interspeech), 2010, pp. 1073–1076.

[17] J.-F. Bonastre, P. Morin, 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), 2003, pp. 2013–2016.

[18] W. Chen, Q. Hong, X. Li, GMM-UBM for text-dependent speaker recognition, in: International Conference on Audio, Language and Image Processing (ICALIP), IEEE, 2012, pp. 432–435.

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

[20] A. Larcher, K. A. Lee, B. Ma, H. T. N. Thuy, Dual scoring for text-dependent speaker verification (2012).

[21] A. E. Rosenberg, C. Lee, S. Gokcen, Connected word talker verification using whole word Hidden Markov Models, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1991, pp. 381–384.

[22] T. Kato, T. Shimizu, Improved speaker, verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003, pp. 57–60.

[23] T. Matsui, S. Furui, Concatenated phoneme models for text-variable speaker recognition, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, 1993, pp. 391–394.

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

[25] M. Schmidt, H. Gish, Speaker identification via support vector classifiers, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE, 1996, pp. 105–108.

[26] C. Dong, Y. Dong, J. Li, H. Wang, Support Vector Machines Based Text Dependent Speaker Verification Using HMM superverctors, in: Speaker and Language Recognition Workshop (Odyssey), 2008, pp. 1–7.

[27] H. Aronowitz, Text-Dependent Speaker Verification Using a Small Development Set, in: Speaker and Language Recognition Workshop (Odyssey), 2012, pp. 1–5.

[28] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li, J.-F. Bonastre, I-vectors in the context of phonetically-constrained short utterances for speaker verification, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, pp. 4773–4776.

[29] D. Sturim, D. Reynolds, R. Dunn, T. Quatieri, Speaker verification using text-constrained Gaussian mixture models, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, IEEE; 1999, 2002, pp. 677–680.

[30] A. Stolcke, A. Mandal, E. Shriberg, Speaker Recognition with Region-Constrained MLLR Transforms, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, pp. 4397–4400.

[31] A. Larcher, J.-F. Bonastre, J. S. D. Mason, Reinforced temporal structure information for embedded utterance-based speaker recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), 2008, pp. 371–374.

[32] A. Larcher, J.-F. Bonastre, J. S. D. Mason, Short utterance-based video aided speaker recognition, in: IEEE International workshop on Multimedia Signal Processing, 2008, pp. 897–901.

[33] A. Larcher, J.-F. Bonastre, J. S. Mason, Constrained Viterbi decoding for embedded user-customised password speaker recognition, in: ACM Symposium On Applied Computing, 2010, pp. 1501–1502.

[34] N. B. Yoma, T. F. Pegoraro, Robust speaker verification with state duration modeling, Speech Communication 38 (1–2) (2002) 77–88.

[35] M. Forsyth, Discriminating observation probability (DOP) HMM for speaker verification, Speech communication 17 (1) (1995) 117–129.

[36] B. Yegnanarayana, S. M. Prasanna, J. M. Zachariah, C. S. Gupta, Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system, IEEE Transactions on Speech and Audio Processing 13 (4) (2005) 575–582.

[37] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, RASTA-PLP speech analysis technique, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, 1992, pp. 121–124.

[38] M. J. Alam, P. Ouellet, P. Kenny, D. O'Shaughnessy, Comparative Evaluation of Feature Normalization Techniques for Speaker Verification, Advances in Nonlinear Speech Processing 1 (2011) 246–253.

[39] D. Wu, B. Li, H. Jiang, Speech recognition, technologies and applications – Normalization and transformation techniques for robust speaker recognition, Artificial intelligence series, InTech, Vienna, Austria, 2008.

[40] D. T. Toledano, C. Esteve-Elizalde, J. Gonzalez-Rodriguez, R. F. Pozo, L. H. Gomez, Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition, in: Speaker and Language Recognition Workshop (Odyssey), Stellenbosch, South Africa, 2008.

[41] B. Dumas, C. Pugin, J. Hennebert, D. Petrovska-Delacrétaz, A. Humm, F. Evéquoz, R. Ingold, D. V. Rotz, MyIdea–Multimodal biometrics database, description of acquisition protocols, Biometrics on the Internet 275 (2005) 59–62.

[42] A. Larcher, J.-F. Bonastre, B. Fauve, K. A. Lee, C. Lévy, H. Li, J. S. Mason, J.-Y. Parfait, ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition, in: Annual Conference of the International Speech Communication Association (Interspeech), 2013, pp. 1–5.

[43] J.-L. Gauvain, C.-H. Lee, Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, 1994, pp. 291–298.

[44] S. J. Young, The general use of tying in phoneme-based HMM speech recognisers, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, 1992, pp. 569–572.

[45] C. Lévy, G. Linares, P. Nocera, J.-F. Bonastre, Mobile phone embedded digit-recognition, Springer Sciences, 2006, Ch. 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2, pp. 71–84.

[46] D. A. Reynolds, R. C. Rose, Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Transactions on Acoustics, Speech and Signal Processing 3 (1) (1995) 72–83.

[47] O. Viikki, K. Laurila, Cepstral domain segmental feature vector normalization for noise robust speech recognition, Speech Communication 25 (1) (1998) 133–147.

[48] P. Nocera, G. Linares, D. Massonié, L. Lefort, Phoneme lattice based A* search algorithm for speech recognition, in: Text, Speech and Dialogue, Springer, 2006, pp. 301–308.

Dr Anthony Larcher:

Anthony is scientist at Institute for Infocomm Research (I2R). He graduated from the Grenoble Institute of Technology, receiving Engineering and M.S. degrees in electronics, electrical engineering, automation and signal processing in 2005. He received the Ph.D. degree in Computer Science from University of Avignon in France in 2009. His current research interests are in the areas of speaker recognition, speech processing and language identification.

Pr Jean-Francois Bonastre:

Jean-Francois Bonastre obtained his Ph.D. degree in 1994 in automatic speaker identification using phonetic-based knowledge. He is full professor at the University of Avignon and a member of the Institut Universitaire de France. He has been vice president of University of Avignon since December 2008. As a member of the Natural Language Processing Group, he developed his research in speaker characterization and recognition using phonetic, statistic and prosodic information, while teaching and lecturing on various subjects covering computer science, speech processing, audio signal classification and indexing, and biometry. In 2002 he took up a one-year sabbatical stay in Panasonic Speech Technology Laboratory, Santa Barbara, CA. From 2001 to 2004, he was the chairman of AFCP, the French-Speaking Speech Communication Association (currently a regional branch of ISCA). He is President of International Speech Communication Association (ISCA) since September 2011, after he has been ISCA vice president since 2007. He was also the chair of the ISCA Speech and Language Characterization SIG for two years. Jean-Francois Bonastre is IEEE Senior Member.

Dr John Mason:

John is currently an Associate Professor at Swansea University. He received MSc and Ph.D. degrees from the University of Surrey in 1971 and 1974 respectively, joining University of Wales Swansea as a lecturer in May 1973. In 1979 he took up a one-year appointment as a senior research engineer at Hewlett Packard Ltd in South Queensferry and in 1994 he was invited to work on an international project at the Australian National University, Canberra, as a visiting research fellow. From the time of his PhD studies through to today his research interest have focused on digital signal processing. Of particular note is the work done on finding solutions to complex Chebyshev approximations, a long-standing problem. Over the last 20 years his research has revolved around speaker recognition. In this area he has served on the technical committees of many international research meetings.