Unsupervised Data-driven Hidden Markov Modeling for Text-dependent Speaker Verification

Dijana Petrovska-Delacrétaz[1] and Houssemeddine Khemiri[2]
[1]*Télécom SudParis, SAMOVAR CNRS, Université Paris-Saclay, Evry, France*
[2]*PW Consultants, TalkToPay, Paris, France*
dijana.petrovska@telecom-sudparis.eu, h.khemiri@pw-consultants.com

Keywords: Unsupervised Data-driven Modeling, Hidden Markov Models, Text-dependent Speaker Verification, Concurrent Scoring.

Abstract: We present a text-dependent speaker verification system based on unsupervised data-driven Hidden Markov Models (HMMs) in order to take into account the temporal information of speech data. The originality of our proposal is to train unsupervised HMMs with only raw speech without transcriptions, that provide pseudo phonetic segmentation of speech data. The proposed text-dependent system is composed of the following steps. First, generic unsupervised HMMs are trained. Then the enrollment speech data for each target speaker is segmented with the generic models, and further processing is done in order to obtain speaker and text adapted HMMs, that will represent each speaker. During the test phase, in order to verify the claimed identity of the speaker, the test speech is segmented with the generic and the speaker dependent HMMs. Finally, two approaches based on log-likelihood ratio and concurrent scoring are proposed to compute the score between the test utterance and the speaker's model. The system is evaluated on Part1 of the RSR2015 database with Equal Error Rate (EER) on the development set, and Half Total Error Rate (HTER) on the evaluation set. An average EER of 1.29% is achieved on the development set, while for the evaluation part the average HTER is equal to 1.32%.

1 INTRODUCTION

The speaker verification task is to decide if a person, who claims to be the target speaker, is or is not that speaker. The decision is either an acceptance or a rejection. Relative to the spoken utterance, speaker verification systems are classified into three categories:

- **text-independent**: the speaker can speak freely during the enrollment and testing phases. A text independent system can recognize a speaker independently of what she/he is saying;
- **text-dependent**: the speaker should reproduce, during the test, the same words or sentences, called pass-phrase, that were pronounced during the enrollment. We are interested in this category. The main challenge of a text-dependent system is to model the speaker characteristics together with the lexical content of the verification utterance.
- **text-prompted**: the speaker should pronounce, during the test, words and sentences proposed by the system. These words or sentences are different from those pronounced during the enrollment phase and can change in each new test.

Text-independent speaker verification, received more attention than the text-dependent task. This could be explained by the international evaluation framework organized by the National Institute of Standards and Technology (NIST) (Martin and Greenberg, 2010). NIST organizes yearly evaluation campaigns for text-independent speaker verification and provides a large amount of speech data, pushing the scientific community to focus more on this task. Moreover, the text-independent scenario corresponds to many applications such as forensic, speaker diarization or speaker tracking.

However, with emerging mobile applications that require identity verification of the speaker, the text-dependent scenario is more appropriate. For such applications, with the assumption of cooperative users, they can be asked to pronounce the same text during both enrollment and test phases. This constraint reduces both the effects of lexical and duration mismatch. Contrary to text-independent speaker verification that requires at least one minute of speech

Petrovska-Delacrétaz, D. and Khemiri, H.
Unsupervised Data-driven Hidden Markov Modeling for Text-dependent Speaker Verification.
DOI: 10.5220/0006202001990207
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 199-207
ISBN: 978-989-758-222-6

to reach high accuracy, text-dependent verification can be done with shorter duration utterances. Reducing the duration and lexical variability improves significantly the performance of text-dependent systems (Hébert, 2008).

Text-dependent speaker verification can be done using Hidden Markov Models (HMMs), dynamic programming or methods adapted from the text-independent systems (Larcher et al., 2014). However, systems based on HMMs are the most common approach for this task. The granularity of the units modeled by the HMM depends on the level of textual transcriptions of the training data. A phonetic-level transcription offers the finest representation of the acoustic space and could be used to train a precise and robust set of HMM models. On the other hand, word- or sentence-level models are limited to a specific lexical content. Moreover, the need of transcribed speech data to develop such systems, could be a major problem, especially for under-resourced languages.

In this paper, a text-dependent speaker verification system based on unsupervised HMM modeling (with no need of transcribed speech data) is proposed. First, a set of generic HMM models is acquired . Then, the acquired models are used to segment the enrollment data of target speakers. Then the enrollment speech data for each target speaker is segmented with the generic models, and further processing is done in order to obtain speaker and text adapted HMMs, that will represent each speaker. During the test phase, in order to verify the claimed identity of the speaker, the test speech is segmented with the generic and the speaker dependent HMMs. Finally, two approaches based on log-likelihood ratio and concurrent scoring are proposed to compute the score between the test utterance and the speaker's model.

The rest of the paper is organized as follows. In Section 2, an overview of text-dependent systems is presented. The proposed approach is described in Section 3. Database, experimental setup and results are given in Section 4. Conclusions and perspectives can be found in Section 5.[1]

2 RELATED RESEARCH

In the last decade, the scientific community focused on text-dependent systems due to its commercial potential. In fact, the fixed-text required by the system and the short duration of enrollment and testing are well suited for commercial applications (Wagner et al., 2006). However, obtaining high accuracy with short enrollment and test utterances, represents scientific and technical challenges.

Text-dependent systems should be able to extract relevant information related to both speaker and lexical content. Three major families for text-dependent speaker verification systems are found in the literature.

The first family is inspired from the text-independent systems. State-of-the-art text-independent systems, based on Gaussian Mixture Models/Universal Background Model (GMM/UBM) and i-vectors, have proven their efficiency. These systems are adapted to take advantages of the lexical information required by text-dependent speaker verification, such as the GMM/UBM system proposed in (Boies et al., 2004), the i-vector system proposed in (Stafylakis et al., 2013) and the joint factor analysis system proposed in (Stafylakis et al., 2016). Other systems based on text-independent scenarios are proposed to model the temporal structure of the speech signal, such as support vector machines (Aronowitz, 2012), artificial neural networks (Woo et al., 2000), and deep neural networks (Variani et al., 2014).

The second family is based on dynamic programming, where speaker and lexical information are extracted at the frame level. In (Furui, 1981), a speaker verification system based on distance computed between cepstrum coefficients of enrollment and testing utterances using dynamic time warping is proposed. In (Dutta, 2008), a system based on spectrogram segmentation and template matching is developed. In addition, vector quantization is combined with dynamic time warping to improve the accuracy of the system in (Bahaghighat et al., 2012). Systems based on dynamic programming are capable of precise modeling of the temporal structure of the pass-phrases. However, the accuracy of these methods is highly affected by the intra-speaker variability. Multiple templates for each of the words of the pass-phrase (Ramasubramanian et al., 2006) can partially sove this problem.

The last family relies on probabilistic approaches where HMMs are exploited to capture the temporal information of the speech signal. HMMs are inherently more robust to intra-speaker variability and allow the modeling of the temporal structure of the speech utterances. Phone-based (Matsui and Furui, 1993), word-based, and sentence-based (Kato and Shimizu, 2003), and (Subramanya et al., 2007) HMM models are proposed in the literature to represent the pass-phrases. Phone-based modeling benefits from the progress achieved in the field of speech recognition and enables precise modeling of the pass-phrases. However phonetic transcriptions are required to ob-

[1] This paper presents work done while Houssemeddine Khemiri was a post-doctoral researcher at Télécom SudParis

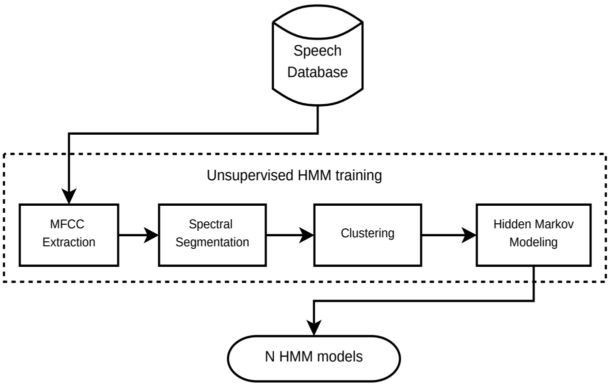

Figure 1: Unsupervised training of HMM models.

tain such models, which are not always available. Word and sentence transcriptions are more easy to acquire, but their HMM models lack generalization power since they are restrained to a limited lexical content.

There are also hybrid systems that combine the GMM/UBM architecture followed by HMM modeling, proposed in (Larcher et al., 2014). This system, referred as HiLAM, consists of a hierarchical three-layer modeling. The first two layers are based on a GMM/UBM approach. The third layer is based on HMM modeling to incorporate the temporal structure information into the speaker model.

In this paper, a pseudo-phone modeling using unsupervised HMM is exploited to develop a text-dependent speaker verification system. The proposed system has the advantage of HMMs, and due to its unsupervised nature does not rely on transcribed speech data. This system is evaluated on Part1 of the RSR2015 database, and is compared to a classical GMM/UBM approach and the HiLAM system.

3 PROPOSED SYSTEM

The proposed system is mainly composed of three steps: unsupervised HMM training, adaptation and scoring. In the first phase, a set of generic HMM models is trained using raw speech data without textual transcriptions. Then an adaptation process is performed to build the target model. In this step speaker-dependent followed by text-dependent adaptation is done. The scoring phase consists on computing a similarity score between a test utterance and the target model. Two similarity scores, based on Log-Likelihood Ratio (*LLR*) and concurrent scoring, are developed.

3.1 Unsupervised HMM Training

Unsupervised HMM training is used for different audio and speech processing fields such as very low bit-rate speech coding (Chollet et al., 1999), text-independent speaker verification (Hannani, 2007), speech recognition (Deligne and Bimbot, 1997), keyword discovery (Siu et al., 2010), topic classification (Siu et al., 2011), and audio indexing (Khemiri et al., 2014).

As shown in Figure 1, the set of data-driven HMM models is automatically acquired through feature extraction, spectral segmentation, clustering, and hidden Markov modeling. More details are available from (Khemiri, 2013). The feature extraction is done with Mel Frequency Cepstral Coefficients (MFCC). Then spectral segmentation is performed to find the stable regions of speech data. These regions represent the spectrally stable segments of the speech data. This process is done by calculating a spectral stability curve obtained by computing the Euclidian distance between two successive feature vectors. The local maxima of this curve represent the segment boundaries while the minima represent the stable parts of the signal. The next step of the training process consists of grouping the obtained segments into *N* classes via vector quantization (Linde et al., 1980). The result of this step is an initial labeling of the training corpus. The final component represents the Hidden Markov modeling, where a set of *N* data-driven HMM units are trained on the basis of the initial segmentation and labeling provided by the previous steps. It is mainly based on Baum-Welch re-estimations (Baum et al., 1970) and on an iterative procedure of refinement of the models. The resulting set of *N* HMMs, referred as generic HMMs, are used to segment any incoming speech data.

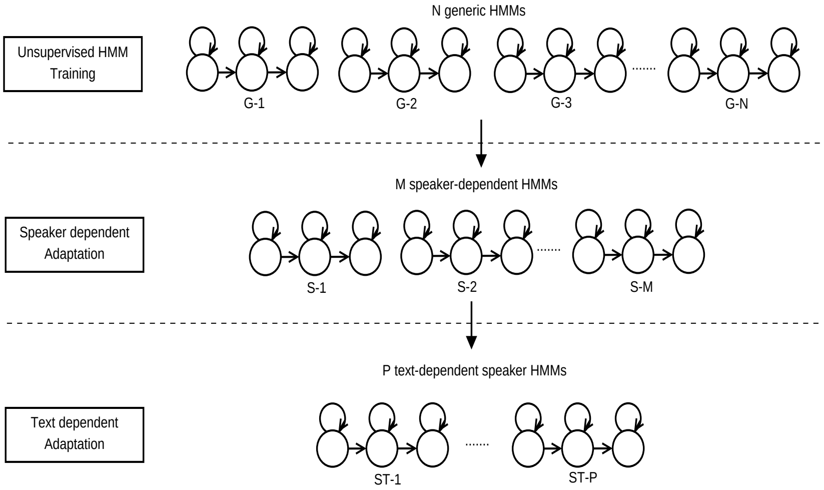
N generic HMMs
Unsupervised HMM Training
G-1
G-2
G-3
G-N
M speaker-dependent HMMs
Speaker dependent Adaptation
S-1
S-2
S-M
P text-dependent speaker HMMs
Text dependent Adaptation
ST-1
ST-P

Figure 2: The adaptation phase in the proposed text-dependent speaker verification system.

3.2 Adaptation

As shown in Figure 2, the adaptation process is performed via speaker- and text-dependent adaptation of the generic models with each client (target) enrollment data, providing the final client specific set of HMM representinf its enrolment pass phrase.

For the speaker-dependent adaptation of the generic HMMs, first the enrollment speech of each speaker is segmented with the Viterbi algorithm (Viterbi, 1967) with the set of generic HMMs. The Viterbi algorithm finds the most likely string of symbols from the set of N Generic HMMs, given the acoustic signal, and the corresponding speech segmentation.[2].Then, only HMM models that are present in this segmentation and represented with enough frames are taken into account in the remaining process. A context independent re-estimation, which adapts each HMM mean component individually, is performed using the Maximum A Posteriori (MAP) criterion. That means that each generic HMM found in the enrollment data will be adapted with all the occurrences of this segment found in the enrollment data. The resulting speaker-dependent HMM models will be used for the text-dependent adaptation.

After the generic HMM models are adapted to the speaker, we continue the adaptation on order to adapt the set of speaker dependent HMM models to the pass phrase of the speaker, in order to have the final speaker and text dependent HMMs, that will represent the speaker enrollment model.This text-dependent adaption is based on iterative adaptation of the speaker-dependent HMM models. First, the enrollment speech is segmented using the speaker-dependent models. Then a context dependent re-estimation which adapts all models in parallel is applied using the MAP criterion and the Viterbi algorithm to obtain a new segmentation using the adapted HMM. This process is iteratively repeated until the HMM models converge or the maximum number of iteration is reached. During this step, the number of HMM models could be reduced by removing those states that are represented by few frames. At the end of this step a set of HMMs, referred as speaker- and text-dependent HMM, is created.

After the enrollment phase, each speaker is represented with a set of speaker-and text-dependent HMMs that are relative to the pass phrase that is chosen for enrollment. During testing, given a speech sequence X, and the enrollment HMMs models of the target we propose two methods of scoring, presented in the next paragraph. Depending on the adopted scoring methods, two target models are considered. For the method based on LLR, the target model is the text-dependent HMM obtained after the iterative adaptation. For the concurrent scoring measure, the text-dependent HMM is grouped with the generic HMM to form the target model.

3.3 Log Likelihood and Concurent Score Computing

Two methods are used to compute a similarity score between a test utterance X and the target speaker model. The first score is based on the well known and

[2] Note that each time we need to decode (segment) the speech with a set of HMMs this is done with the Viterbi algorithm. This algorithm needs an input speech data and a set of HMM models and gives as output a set of symbols corresponding to the most likely HMM models and the corresponding time boundaries

widely used Log-Likelihood Ratio *(LLR)* as follows:

$$S_{LLR}(X) = \log \frac{L_{TD-HMM}(X)}{L_{G-HMM}(X)} \quad (1)$$

where $L_{TD-HMM}(X)$ and $L_{G-HMM}(X)$ are, respectively, the likelihood of the test utterance X given the text-dependent speaker HMM, and the likelihood of X given the the generic HMM.

For the concurrent scoring method, text-dependent speaker HMMs and the generic HMMs are used concurrently in the Viterbi segmentation. The result of this segmentation is a sequence of HMMs symbols that belong either to the text-dependent or to the generic model. A post processing, smoothing procedure is applied to eliminate outliers. This method proposes a voting scheme that uses a sliding window on the HMM sequence. The sliding window operates on an odd number of symbols, and with a majority voting decides if the middle symbol needs to be changed or not. If he middle symbol is the same as the majority vote it is not changed. If the middle symbol is not as the majority vote, then it is change in order to respect the majority voting. For example if the middle symbol belongs to the speaker models and the two left and right neighbors belong to the generic models, then the symbol of the middle is changed as one belonging to the generic models. The final decision of the concurrent scoring score is the ratio between the duration of test segments belonging to the target speaker HMMs and the total duration of the test utterance.

4 EXPERIMENTAL PROTOCOLS AND RESULTS

In this section, the database along with the experimental setups and results are described. The part1 of RSR2015 database (Larcher et al., 2014) is used to evaluate the proposed system. A comparative study is performed with the classical UBM/GMM system and the HiLAM system (Larcher et al., 2014).

4.1 Database and Protocols

The RSR2015 database consists of recording from 157 male and 143 female speakers in 9 sessions using mobile devices. The database is divided into three parts according to the lexical content. Part 1 is dedicated to text-dependent scenario where each speaker records 30 sentences per session selected from the TIMIT database (Garofolo et al., 1993), leading to 72h of audio recordings and approximately 28h of nominal speech. Part2 consists of command control sentences, while Part3 is dedicated to text-prompted speaker verification. Note that only Part1 of the database is used in this paper.

Part1 is divided into three gender-dependent subsets as shown in Table 1. For each speaker of the development and evaluation subsets, 3 sessions are used for enrollment while the remaining sessions are left for tests, leading to mean enrollment durations of 9 seconds.

Table 1: Number of speakers in the background, development and evaluation subsets of Part1 of the RSR2015 database.

Subset	Female	Male
Background	47	50
Development	47	50
Evaluation	49	57

As mentioned before, the text-dependent speaker verification system should be able to decide if the speaker who pronounces the test utterance is the target speaker and if the test utterance matches the enrollment utterance. Therefore, the following four trials are defined:

1. **target-correct (tar-c)**: the target speaker pronounces the expected pass-phrase;
2. **target-wrong (tar-w)**: the target speaker pronounces a wrong pass-phrase (a phrase that is different from the enrollment one);
3. **impostor-correct (imp-c)**: An impostor speaker pronounces the expected pass-phrase;
4. **impostor-wrong (imp-w)**: An impostor speaker pronounces a wrong pass-phrase (a phrase that is different from the enrollment one).

The first trial is the only genuine trial while the others are considered as impostor trials, as defined in (Larcher et al., 2014). The **target-wrong** trials simulate a scenario where and impostor is playing back a recording from the target speaker. The **impostor-correct** trials are more challenging than the **impostor-wrong** ones, as the impostor produces the expected pass-phrase that is used to train the target speaker model. Based on this protocol, the number of trials for each case is reported in Table 2.

4.2 Performance Measure

To measure the performance of the speaker verification systems, two different criteria are used. The Equal Error Rate (EER) is computed for both the development and evaluation parts. The Half Total Error Rate (HTER) is only computed on the evaluation part.

Table 2: Number of trials for each definition on the development and evaluation subsets of Part1 of RSR2015 for female and male.

Trial	Female		Male	
	development	evaluation	development	evaluation
tar-c	8,419	8,631	8,931	10,244
tar-w	244,123	250,299	259,001	297,076
imp-c	387,230	414,249	437,631	573,664
imp-w	5,612,176	6,006,596	6,342,019	8,318,132

To compute the HTER, a threshold θ is defined on the development partition at the EER point. This threshold is applied to the evaluation partition to obtain the HTER as follows:

$$HTER = \frac{FAR(\theta, EVAL) + FRR(\theta, EVAL)}{2} \qquad (2)$$

where FAR is the False Acceptance Rate and FRR is the False Rejection Rate. The HTER measure allows checking whether the threshold defined on the development subset at the EER point is giving a FAR and FRR that are close to the EER on new unseen evaluation data.

4.3 Experimental Settings

The features extraction component is common for the proposed system and the GMM/UBM system. The feature vector is composed of 20 MFCC coefficients together with their first derivatives and the delta energy, leading to a vector with a dimension of 42 features. The speech activity detector is based on three Gaussian's modeling of the energy of the speech data, and is used to determine speech and silence frames. The feature vectors belonging to speech part are normalized to fit a zero-mean and an unit variance distribution.The non-speech feature vectors are used to train a silence HMM model.

Regarding the UBM/GMM system, gender-dependent UBM models are trained with 1024 Gaussians. Then, each target model is created by adapting the mean of the UBM to the enrollment speech data, using the MAP criterion. The score is computed by the log-likelihood ratio between the test feature vector and the target model and the UBM model. Note that only the 10 best Gaussian components are considered for the calculation of the score. The ALIZE_3.0 (Larcher et al., 2013) and SPRO_4.0 (Gravier, 2003) toolkits are used to develop the UBM/GMM system.

For the proposed system, gender-dependent data-driven HMM models are trained on the background subset of the RSR2015 database. Each model is presented by a left-right HMM having three emitting states with no skips. The number of HMM models is empirically fixed to 16 on the development subset of RSR2015. This number is reduced when the speaker- and text-dependent adaptations are applied. The adaptation process is performed on the mean parameters of HMMs using the HTK toolkit (Young et al., 2006). For the system based on concurrent scoring the size of the smoothing window is equal to 5.

4.4 Results

In order to exploit the temporal information of speech data, we propose to use data-driven HMM training, with no need of transcribed speech data. For this purpose, two systems based on LLR scoring and concurrent scoring are developed. All the experiences are done following the experimental protocol explained in (Larcher et al., 2014). Table 3 and 4 show, respectively, the EER on the development subset and the EER, and HTER on the evaluation subsets for the two systems. Note that the 90% confidence interval of the EER varies between 0.17% and 0.22%.

The system based on concurrent scoring outperforms the LLR-based method for all trials for development and evaluation subsets. This is explained by the post-processing performed on the segmentation provided by the combined data-driven HMM models to eliminate the outliers that could occur in that segmentation. In fact, the absolute improvement achieved in terms of EER by introducing the smoothing post-processing varies between 1% and 2%. On the other hand, the threshold fixed on the development set at the EER point, does not provide the optimal performances on the evaluation subset. In fact the difference between the EER and HTER could reach 2.5%. This could be explained by the mismatch between the development and evaluation subsets. In addition, the results show that for both methods, the system can discriminate better a target speaker pronouncing a wrong sentence than an impostor who pronounces the correct pass-phrase. This shows that the acoustic characteristics related to the lexical content are more represented in the text-dependent speaker model than those related to the speaker.

The next step is to compare the best system based

Table 3: Performance of the data-driven HMM systems based on the LLR and the concurrent scoring on the **development** subset in terms of EER for the **target-correct (tar-c)**, **target-wrong (tar-w)**, **impostor-correct (imp-c)** and **impostor-wrong (imp-w)** trials.

	Female		**Male**	
	LLR	**Concurrent Scoring**	**LLR**	**Concurrent scoring**
tar-c/tar-w	2.48	1.25	3.12	1.63
tar-c/imp-c	3.58	1.54	4.28	2.26
tar-c/imp-w	1.22	0.52	1.53	0.55

Table 4: Performance of the data-driven HMM systems based on the LLR and the concurrent scoring on the **evaluation** subset in terms of EER, and HTER for the **target-correct (tar-c)**, **target-wrong (tar-w)**, **impostor-correct (imp-c)** and **impostor-wrong (imp-w)** trials.

	Female				**Male**			
	LLR		**Concurrent Scoring**		**LLR**		**Concurrent Scoring**	
	EER	HTER	EER	HTER	EER	HTER	EER	HTER
tar-c/tar-w	1.27	1.77	0.77	1.58	1.04	1.11	0.81	0.9
tar-c/imp-c	2.53	4.95	1.16	2.15	3.58	6.37	1.66	2.36
tar-c/imp-w	0.98	1.56	0.27	0.48	0.88	1.23	0.20	0.45

Table 5: Performance of proposed system, the UBM/GMM and HiLAM systems (Larcher et al., 2014) on the **development** subset in terms of EER for the **target-correct (tar-c)**, **target-wrong (tar-w)**, **impostor-correct (imp-c)** and **impostor-wrong (imp-w)** trials.

	Female			**Male**		
	UBM/GMM	**HiLAM**	**Proposed System**	**UBM/GMM**	**HiLAM**	**Proposed System**
tar-c/tar-w	2.31	1.77	1.25	3.17	1.66	1.63
tar-c/imp-c	3.00	3.24	1.54	3.59	3.69	2.26
tar-c/imp-w	0.33	0.45	0.52	0.62	0.49	0.55

Table 6: Performance of proposed system, the UBM/GMM and HiLAM systems (Larcher et al., 2014) on the **evaluation** subset in terms of EER for the **target-correct (tar-c)**, **target-wrong (tar-w)**, **impostor-correct (imp-c)** and **impostor-wrong (imp-w)** trials.

	Female			**Male**		
	UBM/GMM	**HiLAM**	**Proposed System**	**UBM/GMM**	**HiLAM**	**Proposed System**
tar-c/tar-w	1.59	0.61	0.77	2.21	0.82	0.81
tar-c/imp-c	1.55	2.96	1.16	2.16	2.47	1.66
tar-c/imp-w	0.20	0.14	0.27	0.28	0.19	0.20

on concurrent scoring with a baseline GMM/UBM system, that we implemented, and with the published results related to the HiLAM system (Larcher et al., 2014). Table 5 and 6 give the results of the data-driven HMM system using the concurrent scoring method, the UBM/GMM and HiLAM systems on the development and evaluation subsets, in terms of EER. These results show that the classical UBM/GMM system gives better results than the HiLAM system in the case where the impostor pronounces the correct pass-phrase. In addition, the proposed system outperforms both UBM/GMM and HiLAM systems for the majority of trials. In fact, a significant improvement, reaching 1.5% in terms of EER, could be seen especially in the case of **tar-c/imp-c**. It is important to note that the HiLAM system, as presented in (Larcher et al., 2014), exploits all the enrollment data (30 X 3 sentences) to obtain the text-independent speaker model. While for the data-driven HMM system only the enrolment sequences related to the pass-phrase (3 sentences) are exploited. These results show the efficiency of the proposed system to embed the temporal information of the pass-phrase, even if the amount of enrollment data is limited. Furthermore, since only the HMMs that are significantly represented in the pass-phrase are kept to model the speaker and lexicon information, the text-dependent speaker model is more robust and provides a precise modeling of the temporal structure of the pass-phrase.

Regarding the computational time, the

UBM/GMM system needs 0.03 second to process 1 second of data in the adaptation phase and 0.01 second in the scoring phase. While for the proposed HMM system, the time needed for the adaptation to process 1 second of data is 1.2 second. For the scoring part, the LLR and the concurrent scoring methods require, respectively, 0.07 second and 0.6 second to process 1 second of data.

5 CONCLUSIONS AND PERSPECTIVES

In this paper, a data-driven HMM modeling is proposed for text-dependent speaker verification to exploit the temporal information of speech data. The data-driven models are trained on raw speech data to obtain a set of generic HMMs. This set is then adapted to the target speaker and lexical content of the pass-phrase. Two systems based on log-likelihood radio and concurrent scoring are introduced. The systems are evaluated on Part1 of RSR2015 database. This evaluation shows that concurrent scoring system is more accurate than the one based on the log-likelihood ratio. Moreover, the results show the relevance of the proposed method when compared with an UBM/GMM and the HiLAM systems. Future works will be dedicated on the evaluation of the proposed system on Part2 and 3 of the RSR2015 database. In addition, the concurrent scoring method should be accelerated in case of integration on a mobile device.

REFERENCES

Aronowitz, H. (2012). Text dependent speaker verification using a small development set. In *The IEEE Odyssey Speaker and Language Recognition Workshop*.

Bahaghighat, M. K., Sahba, F., and Tehrani, E. (2012). Text-dependent speaker recognition by combination of lbg vq and dtw for persian language. *International Journal of Computer Applications*, 51(16):23–27.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

Boies, D., Hébert, M., and Heck, L. (2004). Study on the effect of lexical mismatch in text-dependent speaker verification. In *The IEEE Odyssey Speaker and Language Recognition Workshop*, pages 1–5.

Chollet, G., Černocký, J., Constantinescu, A., Deligne, S., and Bimbot, F. (1999). *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*, pages 375–388. NATO ASI Series. Springer Verlag.

Deligne, S. and Bimbot, F. (1997). Inference of variable-length linguistic and acoustic units by multigrams. *Speech Communication*, 23(3):223–241.

Dutta, T. (2008). Dynamic time warping based approach to text-dependent speaker identification using spectrograms. In *Congress on Image and Signal Processing*, volume 2, pages 354–360.

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

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (1993). Timit acoustic-phonetic continuous speech corpus. In *Linguistic Data Consortium*.

Gravier, G. (2003). *Speech Signal Processing Toolkit, release 4.0*.

Hannani, A. E. (2007). *Text-Independant Speaker Verification Based On High-Level Information Extracted With Data-Driven Methods*. PhD thesis, University of Fribourg (Switzerland) and INT/SITEVRY (France).

Hébert, M. (2008). Text-dependent speaker recognition. In *Springer handbook of speech processing*, pages 743–762. Springer.

Kato, T. and Shimizu, T. (2003). 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*, volume 2, pages 57–60.

Khemiri, H. (2013). *Unified data-driven approach for audio indexing, retrieval and recognition*. Theses, Télécom ParisTech.

Khemiri, H., Petrovska-Delacrétaz, D., and Chollet, G. (2014). Alisp-based data compression for generic audio indexing. In *Data Compression Conference*, pages 273–282.

Larcher, A., Bonastre, J., Fauve, B., Lee, K., Lévy, C., Mason, H. L. J., and Parfait, J. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *the Annual Conference of the International Speech Communication Association (Interpseech)*, pages 2768–2773.

Larcher, A., Lee, K., Ma, B., and Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Communication*, 60:56 – 77.

Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95.

Martin, A. F. and Greenberg, C. S. (2010). The NIST 2010 speaker recognition evaluation. In *the Annual Conference of the International Speech Communication Association (Interpseech)*, pages 2726–2729.

Matsui, T. and Furui, S. (1993). Concatenated phoneme models for text-variable speaker recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 391–394.

Ramasubramanian, V., Das, A., and Kumar, V. P. (2006). Text-dependent speaker-recognition using one-pass dynamic programming algorithm. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–I.

Siu, M., Gish, H., Chan, A., and Belfield, W. (2010). Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision. In *the Annual Conference of the International Speech Communication Association (Interpseech)*.

Siu, M., Gish, H., Lowe, S., and Chan, A. (2011). Unsupervised audio pattern discovery using hmm-based self-organized units. In *the Annual Conference of the International Speech Communication Association (Interpseech)*.

Stafylakis, T., Kenny, P., Alam, M., and Kockmann, M. (2016). Speaker and channel factors in text-dependent speaker recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24:65–78.

Stafylakis, T., Kenny, P., Ouellet, P., Perez, J., Kockmann, M., and Dumouchel, P. (2013). Text-dependent speaker recognition using plda with uncertainty propagation. In *the Annual Conference of the International Speech Communication Association (Interpseech)*, page 36843688.

Subramanya, A., Zhang, Z., Surendran, A., Nguyen, P., Narasimhan, M., and Acero, A. (2007). A generative-discriminative framework using ensemble methods for text-dependent speaker verification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 225–228.

Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4080–4084.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

Wagner, M., Summerfield, C., Dunstone, T., Summerfield, R., and Moss, J. (2006). An evaluation of ”commercial off-the-shelf” speaker verification systems. In *The IEEE Odyssey Speaker and Language Recognition Workshop*, pages 1–8.

Woo, S., Lim, C., and Osman, R. (2000). Text-dependent speaker recognition using the fuzzy artmap neural network. In *IEEE International Conference on Electrical and Electronic Technology*, volume 1, pages 33–38.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*.