MODELLING THE ALTERNATIVE HYPOTHESIS FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore

{*alarcher,kalee,mabin,hli*}*@i2r.a-star.edu.sg*

ABSTRACT

This paper describes text-dependent speaker verification as a task involving four classes of trials depending on whether the target speaker or an impostor pronounces the expected pass-phrase or not. These four classes are used to reformulate the log-likelihood ratio traditionally used in text-independent speaker verification. Three formulations of the alternative hypothesis are considered, leading to three new expressions of the verification score. Experiments performed on the publicly available RSR2015 database show a significant improvement compared to existing baseline scores. A relative gain up to 61% in term of minimum cost is achieved when considering that the alternative hypothesis is the union of three sub-hypotheses corresponding to the three existing classes of impostures.

Index Terms— Speaker verification, Text-Dependent, Impostures

1. INTRODUCTION

Authentication of a person can rely on three types of information: a possession, a knowledge or a biometric sample [1]. In text-dependent speaker verification, a specific case of speaker recognition, an automatic system is expected to only authorize access to a person who can match both the voice characteristic and lexical content. The advantage is twofold. First, security is strengthened by verifying both the knowledge and biometric sample. Second, constraining the lexical content of the spoken utterance improves the performance of speaker verification systems when dealing with short duration speech segments [2, 3]. There are different ways to constrain the lexical content [4] and we focus this work on the case where each target speaker is free to choose from a finite set of personal pass-phrases.

In general, speaker verification[5] is a binary classification task. Given a verification trial involving a target speaker $\mathcal{X}$ and a speech segment $\mathcal{O}$, an automatic system has to decide whether the hypothesis, $H_{\mathcal{X}}$, that the speech segment was spoken by the target speaker is true or not. Ideally, a verification score, reflecting the confidence of the system in hypothesis $H_{\mathcal{X}}$, is computed as a log-likelihood ratio between $H_{\mathcal{X}}$ and its alternative $H_{\overline{\mathcal{X}}}$, for which the speech segment was spoken by an impostor [6].

When introducing text-dependency, the recognition task becomes two-dimensional as the system has to determine if the speech segment was spoken by the target speaker or by an impostor but also if the lexical content, i.e. the pass-phrase $\mathcal{P}$, is correct. Speaker recognition systems are now exposed to four classes of trials described in Table 1. A new verification hypothesis has to be considered: $H_{(\mathcal{X},\mathcal{P})}$ for which the speech segment is the correct pass-phrase spoken by the target speaker. In other words, hypothesis $H_{(\mathcal{X},\mathcal{P})}$ states that the trial belongs to the class $(\mathcal{X},\mathcal{P})$. Subsequently, the new alternative hypothesis $H_{\overline{(\mathcal{X},\mathcal{P})}}$ can be defined as the union of three exclusive sub-hypotheses; namely, $H_{(\mathcal{X},\overline{\mathcal{P}})}$ in which the target speaker pronounces a wrong pass-phrase, $H_{(\overline{\mathcal{X}},\mathcal{P})}$ in which an impostor pronounces the correct pass-phrase and $H_{(\overline{\mathcal{X}},\overline{\mathcal{P}})}$ in which an impostor pronounces a wrong pass-phrase.

Table 1: Four classes of trials existing for text-dependent speaker verification task.

	Correct Pass-Phrase $\mathcal{P}$	Wrong Pass-Phrase $\overline{\mathcal{P}}$
Target Speaker $\mathcal{X}$	$(\mathcal{X},\mathcal{P})$	$(\mathcal{X},\overline{\mathcal{P}})$
Impostor $\overline{\mathcal{X}}$	$(\overline{\mathcal{X}},\mathcal{P})$	$(\overline{\mathcal{X}},\overline{\mathcal{P}})$

Approaches existing in the literature propose to solve this problem by applying a two steps process [7, 8]. First a speech recognition system verifies the pass-phrase, then a text-independent speaker verification system tests the identity of the speaker. If both pass-phrase and identity are correct, the trial is accepted. However, running two systems in parallel increases the demand of computing resources that may become critical for specific applications. In this work we propose to base the verification decision on a single log-likelihood ratio which considers the composite nature of the alternative hypothesis. Indeed, recent works [9, 10] on text-independent speaker verification have shown that an appropriate definition and modelling of the alternative hypothesis improves the performance of the speaker verification system. A single system, introduced in [11] and based on a hierarchical acoustic model, is used to model the verification hypothesis, $H_{(\mathcal{X},\mathcal{P})}$, as well as its alternative $H_{\overline{(\mathcal{X},\mathcal{P})}}$.

In the following section, we introduce the task of text-dependent speaker verification and propose different approaches to model the alternative hypothesis in order to form a log-likelihood ratio. Section 3 describes the acoustic architecture that is used to model the four hypothesis and approximate the verification scores. The performance of the different scoring methods are then compared in Section 4. Finally we discuss the benefit of this approach and the possible extensions in Section 5.

2. TEXT-DEPENDENT SPEAKER VERIFICATION

2.1. Log-Likelihood Ratio Test

In text-independent speaker verification, given an utterance $\mathcal{O}$ and a speaker $\mathcal{X}$, answering the verification task consists of testing the hypothesis $H_{\mathcal{X}}$ that $\mathcal{O}$ was spoken by $\mathcal{X}$ against its alternative hypothesis, $H_{\overline{\mathcal{X}}}$, that $\mathcal{O}$ has not been spoken by $\mathcal{X}$. The decision can

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

be obtained by comparing a log-likelihood ratio between the probability distributions of both hypotheses, $p(\mathcal{O}|H_{\mathcal{X}})$ and $p(\mathcal{O}|H_{\overline{\mathcal{X}}})$, to a fixed threshold Θ, as follows

$$\log p(\mathcal{O}|H_{\mathcal{X}}) - \log p(\mathcal{O}|H_{\overline{\mathcal{X}}}) \lessgtr \Theta \begin{cases} H_{\mathcal{X}} \text{ rejected} \\ H_{\mathcal{X}} \text{ accepted} \end{cases} \quad (1)$$

In addition to the speaker identity, a text-dependent speaker verification system also needs to verify that the speaker pronounces a given pass-phrase $\mathcal{P}$. Thus, the null hypothesis to test, $H_{(\mathcal{X},\mathcal{P})}$, assumes that $\mathcal{O}$ belongs to the class $(\mathcal{X},\mathcal{P})$, for which the pass-phrase $\mathcal{P}$ is spoken by $\mathcal{X}$. A new composite alternative hypothesis, $H_{\overline{(\mathcal{X},\mathcal{P})}}$, is then defined accordingly by considering that the absolute complement of the class $(\mathcal{X},\mathcal{P})$ is the union of three exclusive classes such that:

$$P(\mathcal{O}|H_{\overline{(\mathcal{X},\mathcal{P})}}) = P(\mathcal{O}|H_{(\mathcal{X},\overline{\mathcal{P}})}) + P(\mathcal{O}|H_{(\overline{\mathcal{X}},\mathcal{P})}) + P(\mathcal{O}|H_{(\overline{\mathcal{X}},\overline{\mathcal{P}})}) \quad (2)$$

These three classes are: $(\mathcal{X},\overline{\mathcal{P}})$, in which $\mathcal{O}$ is the speaker $\mathcal{X}$ pronouncing a pass-phrase different from $\mathcal{P}$, $(\overline{\mathcal{X}},\mathcal{P})$ where $\mathcal{O}$ is a speaker different from $\mathcal{X}$ pronouncing the pass-phrase $\mathcal{P}$ and $(\overline{\mathcal{X}},\overline{\mathcal{P}})$ where $\mathcal{O}$ is a speaker different from $\mathcal{X}$ pronouncing a pass-phrase different from $\mathcal{P}$. Note that these classes correspond to the three classes of trials shadowed in Table 1.

2.2. The Case of a Composite Alternative Hypothesis

Due to its composite nature, it could be difficult to estimate the probability distributions of the alternative hypothesis. One way to alleviate this difficulty is to consider separately each class of trials and to compute the probability distribution, $p(\mathcal{O}|H_{\overline{(\mathcal{X},\mathcal{P})}})$ as a combination of the likelihoods of the three sub-hypotheses, $H_{(\mathcal{X},\overline{\mathcal{P}})}$, $H_{(\overline{\mathcal{X}},\mathcal{P})}$ and $H_{(\overline{\mathcal{X}},\overline{\mathcal{P}})}$, that compose $H_{\overline{(\mathcal{X},\mathcal{P})}}$.

The first approach we propose to compute $H_{\overline{(\mathcal{X},\mathcal{P})}}$ is commonly used in speech recognition [12] and language detection [13, 14]. It could be related to a weighted mean of the likelihood of the competing sub-hypotheses given by:

$$p(\mathcal{O}|H_{\overline{(\mathcal{X},\mathcal{P})}}) = \left(\frac{1}{N} \sum_{c \, \in \, \Omega} p(\mathcal{O}|H_c)^{\eta} \right)^{\frac{1}{\eta}} \quad (3)$$

Where Ω is a set of N trial classes that form $H_{\overline{(\mathcal{X},\mathcal{P})}}$ and η is a positive constant. Here, $\Omega = \{(\mathcal{X},\overline{\mathcal{P}}); (\overline{\mathcal{X}},\mathcal{P}); (\overline{\mathcal{X}},\overline{\mathcal{P}})\}$.

The second approach proposed here, is equivalent to the fusion of score which is used to combine several systems in speaker verification [15]. We compute now the logarithm of $p(\mathcal{O}|H_{\overline{(\mathcal{X},\mathcal{P})}})$ as the mean of the log-likelihood of the competing sub-hypotheses such that:

$$\log p(\mathcal{O}|H_{\overline{(\mathcal{X},\mathcal{P})}}) = \frac{1}{N} \sum_{c \, \in \, \Omega} \log p(\mathcal{O}|H_c) \quad (4)$$

In practice, it is not possible to estimate precisely $p(\mathcal{O}|H_{(\mathcal{X},\mathcal{P})})$ and $p(\mathcal{O}|H_{\overline{(\mathcal{X},\mathcal{P})}})$. These likelihood are thus approximated by scores $s(\mathcal{O}|\lambda_{(\mathcal{X},\mathcal{P})})$ and $s(\mathcal{O}|\lambda_{\overline{(\mathcal{X},\mathcal{P})}})$ where $\lambda_{(\mathcal{X},\mathcal{P})}$ and $\lambda_{\overline{(\mathcal{X},\mathcal{P})}}$ are statistical model representing hypotheses $H_{(\mathcal{X},\mathcal{P})}$ and $H_{\overline{(\mathcal{X},\mathcal{P})}}$ respectively.

3. MODELLING OF THE FOUR HYPOTHESES

Fair comparison of several hypotheses requires consistency across hypotheses modelling. For this reason, we use in this work the hierarchical multi-layer acoustic model (HiLAM) introduced in [11, 3]. Based on this acoustic model, which is an extension of the well known GMM/UBM paradigm [16], we propose an approximation of the verification scores described above.

3.1. The Hierarchical Multi-Layer Acoustic Model (HiLAM)

HiLAM is a three layer acoustic architecture described in Figure 1. The upper layer model, λ_{ubm}, is a universal background model (UBM) trained on a reasonably large amount of data. If the training data is covering sufficiently large number of speakers and lexical content, λ_{ubm} is considered speaker- and text-independent. Similarly to the text-independent scenario [16], the UBM is used in the rest of the paper to model the hypothesis $H_{\overline{\mathcal{X}}}$ such that $P(\mathcal{O}|H_{\overline{\mathcal{X}}}) = P(\mathcal{O}|H_{(\overline{\mathcal{X}},\mathcal{P})}) + P(\mathcal{O}|H_{(\overline{\mathcal{X}},\overline{\mathcal{P}})})$

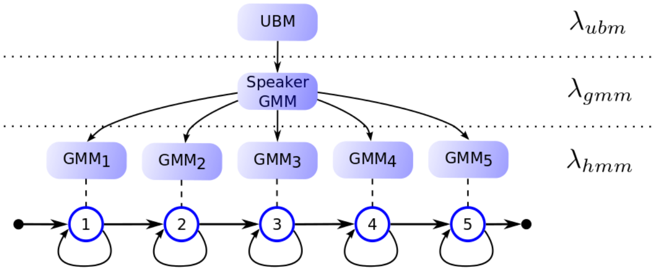

Fig. 1: The hierarchical multi-layer acoustic model (HiLAM)

The middle layer of the HiLAM is a speaker-dependent Gaussian mixture model (GMM). This model, λ_{gmm}, is adapted from the UBM using all data available from the target speaker and a Maximum A Posteriori (MAP) criteria [17]. If the lexical content pronounced by the target speaker during the enrolment is large enough, λ_{gmm} could be considered text-independent. However, we propose to use λ_{gmm} to represent the hypothesis $H_{(\mathcal{X},\overline{\mathcal{P}})}$, assuming that when the enrolment material covers a large lexical content, then $H_{(\mathcal{X},\overline{\mathcal{P}})}$ tends to the text-independent hypothesis $H_{\mathcal{X}}$ if not using discriminative training.

The bottom layer of the HiLAM is a hidden Markov model (HMM) modelling a specific pass-phrase. Each state of this HMM is a GMM derived from the speaker-dependent, text-independent GMM from the second layer, λ_{gmm}, by using all recordings of the given pass-phrase from the target speaker enrolment. This HMM, referred to as λ_{hmm}, is a speaker- and text-dependent model that is used to represent the hypothesis $H_{(\mathcal{X},\mathcal{P})}$. More details about the HiLAM or its training process can be find in [11, 18].

3.2. Score Approximation

During the authentication phase, the likelihood of a given test utterance $\mathcal{O}$ is computed over each of the thee layers of the HiLAM architecture:

- $\Lambda(\mathcal{O}|\lambda_{ubm})$ is the likelihood of utterance $\mathcal{O}$ against the speaker- and text-independent first layer of the HiLAM,
- $\Lambda(\mathcal{O}|\lambda_{gmm})$ is the likelihood of utterance $\mathcal{O}$ against the speaker-dependent, text-independent middle layer of the HiLAM,
- $\Lambda(\mathcal{O}|\lambda_{hmm})$ is the likelihood of utterance $\mathcal{O}$ aligned on the speaker- and text-dependent HMM by using a Viterbi decoding. Probability of transition are not considered here [18].

Those likelihoods can then be used to compute the verification score as a log-likelihood ratio in which the alternative hypothesis is approximated according to the expressions given in Section 2.2. Note that by modelling the four hypotheses with the HiLAM architecture, we assume that λ_{ubm} models $H_{(\overline{\mathcal{X}},\mathcal{P})} \cup H_{(\overline{\mathcal{X}},\overline{\mathcal{P}})}$. Thus, the number of sub-hypotheses of $H_{\overline{(\mathcal{X},\mathcal{P})}}$, N, is reduced to 2 and $\Omega = \{(\mathcal{X},\overline{\mathcal{P}});\overline{\mathcal{X}}\}$.

A first score, $\mathcal{S}_1^{\eta}(\mathcal{O})$, is adapted from Equation 3.

$$\mathcal{S}_1^{\eta}(\mathcal{O}) = \log \Lambda(\mathcal{O}|\lambda_{hmm}) - \log \left[\left(\frac{\Lambda(\mathcal{O}|\lambda_{gmm})^{\eta}}{2} + \frac{\Lambda(\mathcal{O}|\lambda_{ubm})^{\eta}}{2} \right)^{\frac{1}{\eta}} \right] \tag{5}$$

When η tends to infinity, $\mathcal{S}_1^{\eta}(\mathcal{O})$ tends to $\mathcal{S}_1^{max}(\mathcal{O})$ given by:

$$\mathcal{S}_1^{max}(\mathcal{O}) = \log \Lambda(\mathcal{O}|\lambda_{hmm}) - \log \max \left\{ \Lambda(\mathcal{O}|\lambda_{gmm}), \Lambda(\mathcal{O}|\lambda_{ubm}) \right\} \tag{6}$$

The expression of the alternative hypothesis from Eq.4 leads to a third expression of the verification score, $\mathcal{S}_2(\mathcal{O})$, given by:

$$\mathcal{S}_2(\mathcal{O}) = \log \Lambda(\mathcal{O}|\lambda_{hmm}) - \left[\frac{\log \Lambda(\mathcal{O}|\lambda_{gmm})}{2} + \frac{\log \Lambda(\mathcal{O}|\lambda_{ubm})}{2} \right] \tag{7}$$

Two other verification scores are given here for comparison. The baseline, $\mathcal{S}_{HMM}(\mathcal{O})$, is the natural text-dependent score computed from the HiLAM architecture and used in [19, 20].

$$\mathcal{S}_{HMM}(\mathcal{O}) = \log \Lambda(\mathcal{O}|\lambda_{hmm}) - \log \Lambda(\mathcal{O}|\lambda_{ubm}) \tag{8}$$

In this expression, the alternative hypothesis is only modelled by the first layer of the HiLAM architecture, λ_{ubm}, that can be seen as a rough approximation of the $H_{\overline{(\mathcal{X},\mathcal{P})}}$ hypothesis. Eventually, the classic GMM/UBM score, $\mathcal{S}_{GMM}(\mathcal{O})$, is considered for comparison.

$$\mathcal{S}_{GMM}(\mathcal{O}) = \log \Lambda(\mathcal{O}|\lambda_{gmm}) - \log \Lambda(\mathcal{O}|\lambda_{ubm}) \tag{9}$$

4. EXPERIMENTS

4.1. Performance Estimator

Experiments have been conducted to compare the performance of the different scoring methods when dealing with the different classes of impostures existing in text-dependent speaker verification. Amongst the three classes of impostures, the case of an impostor pronouncing a wrong lexical content $(\overline{\mathcal{X}},\overline{\mathcal{P}})$, is the easiest to reject as neither identity nor the lexical content is correct [3]. This class of imposture does not represent a major threat to the system and would make the result look over optimistic. Therefore, we report the performance in terms of minimum detection cost where we exclude the case of an impostor pronouncing a wrong lexical content. This cost function, similar to the one used for the NIST-SRE 2012 evaluation [21], is a single estimator that takes into account the two classes of impostures $(\mathcal{X},\overline{\mathcal{P}})$ and $(\overline{\mathcal{X}},\mathcal{P})$. The cost function is given by

$$C_{Norm} = P_{Miss|\mathcal{X},\mathcal{P}} + \beta \times (0.5 \times P_{FA|\mathcal{X},\overline{\mathcal{P}}}) \tag{10}$$

with $\beta = \frac{C_{FA}}{C_{Miss}} \times \frac{(1-P_{\mathcal{X},\mathcal{P}})}{P_{\mathcal{X},\mathcal{P}}}$ where the parameters of this function are:

- $P_{\mathcal{X},\mathcal{P}}$, the a priori probability that the test speaker is the target speaker,
- $P_{Miss|\mathcal{X},\mathcal{P}}$, the miss error probability,
- $P_{FA|\mathcal{X},\overline{\mathcal{P}}}$, the false alarm error probability for target speaker pronouncing a wrong pass-phrase,
- $P_{FA|\overline{\mathcal{X}},\mathcal{P}}$, the false alarm error probability for impostor pronouncing the correct pass-phrase,
- C_{FA}, the cost of a false alarm,
- C_{Miss}, the cost of a miss.

Note that the *a priori* probability of an impostor trial to be of class $(\mathcal{X},\overline{\mathcal{P}})$ or $(\overline{\mathcal{X}},\mathcal{P})$ are considered equal and that the costs C_{FA} and C_{Miss} are set to 1. Eventually, two values of Cnorm will be reported when setting the *a priori* probability of target speaker pronouncing the correct pass-phrase to different values:

$$\begin{cases} \text{Cnorm}_A & \text{for } P_{\mathcal{X},\mathcal{P}} = 0.01 \\ \text{Cnorm}_B & \text{for } P_{\mathcal{X},\mathcal{P}} = 0.001 \end{cases} \tag{11}$$

4.2. Experimental Protocol

Experiments are conducted on 50 speakers of the Part 1 of the RSR2015 database [19, 20]. In this corpus, each speaker recorded 9 sessions using several mobile devices. During each session, all speakers read a common set of 30 short sentences ($<$ 3seconds) that are used as pass-phrases for the regarded text-dependent speaker verification task. Out of the 9 recording sessions, 3 are used for enrolment while 6 are set aside to be used as testing material. For each speaker, the 30 pass-phrases of the 3 enrolment sessions are used to adapt a text-independent GMM (second layer of the HiLAM architecture). Each text-independent GMM is then used to adapt one speaker- and text-dependent HMMs for each of the 30 pass-phrases by using the 3 sessions of this specific pass-phrase (third layer of the HiLAM). Through this process, 1,500 models are generated (50 speakers, 30 pass-phrases).

The UBM is trained on 6,435 utterances from 50 different male speakers, using materials from the forthcoming Parts 2 and 3 of the RSR2015 database. Thus, none of the 30 pass-phrases has been seen by the UBM.

During the testing phase, for a given speaker, each of the 30 models is compared to all pass-phrases from the remaining 6 sessions of this speaker, generating both $(\mathcal{X},\mathcal{P})$ and $(\mathcal{X},\overline{\mathcal{P}})$ trials. Trials involving impostor speakers are generated by testing all models from the selected speaker against the test material from the 49 remaining speakers of the test-set. Note that in order to limit the number of impostor trials for which the the speaker pronounces a wrong pass-phrase, $(\overline{\mathcal{X}},\overline{\mathcal{P}})$, each model is tested against a sub-set of test segment randomly chosen to cover all speakers, sessions and pass-phrases of the test-set. The number of trials resulting from this process is given in Table 2

Table 2: Number of tests performed for each of the four classes of trials existing for text-dependent speaker verification.

Trial definition	Speaker	Pass-Phrase	Number of tests
$(\mathcal{X},\mathcal{P})$	target-speaker	correct	8,931
$(\mathcal{X},\overline{\mathcal{P}})$	target-speaker	wrong	259,001
$(\overline{\mathcal{X}},\mathcal{P})$	impostor	correct	437,631
$(\overline{\mathcal{X}},\overline{\mathcal{P}})$	impostor	wrong	6,342,019

4.3. System Configuration

Front-end processing produced 50 dimensions acoustic features (19 MFCC, 19 derivatives, first 11 second derivatives and the delta energy). Acoustic features are computed on a 20ms sliding window with shifting of 10ms. Feature of lower energy are discarded and mean variance normalization is applied. Each node of the HiLAM architecture is a 64-distribution GMM. Speaker- and Text-dependent models are 5 states HMMs.

4.4. Performance Analysis

First we evaluate the effect of the parameter η in the expression of the score $\mathcal{S}_1^\eta$. Figure 2 show the evolution of the two minimum costs as a function of η. This experiment shows that the best performance is obtained when $\eta = 0.1$ and that increasing the value of this parameter degrades the performance. For the next experiment, η is fixed to 0.1.

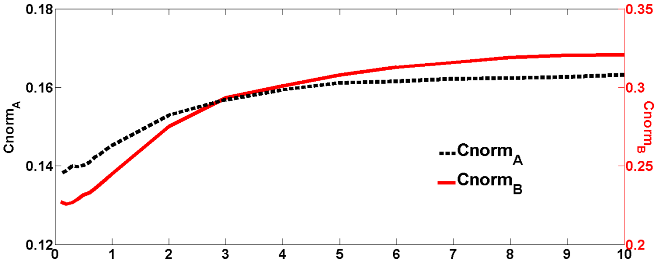

Fig. 2: Variation of the minimum costs for different values of the η parameter when computing the $\mathcal{S}_1^\eta$ score.

Table 3 contains the performance of the five scoring methods proposed in Section 3.2, in terms of minimum cost. A first look at the result shows that S_1^η obtains the lower cost for both functions but that the difference with S_2 is not significant. It is clear that the group of scores including S_1^η, S_1^{max} and S_2, which use a composite modelling of the alternative hypothesis $H_{\overline{(\mathcal{X},\mathcal{P})}}$ outperform the two scores, S_{HMM} and S_{GMM} that use a classical UBM model. Compared to S_{HMM}, S_1^η reduces Cnorm$_A$ and Cnorm$_B$ minimum costs by 61% and 48% respectively. As expected, S_1^{max} does not perform as well as S_1^η. Indeed, S_1^{max} is the limit of S_1^η when η tends to infinity and Figure 2 shows that performance of the system degrades when increasing η. Finally, S_{GMM} obtains maximal value of 1 for each of the minimum cost Cnorm$_A$ and Cnorm$_B$. This result is explained below by analyzing Table 4.

Table 3: Performances of different scoring methods given as minimum detection cost for two values of P_{target}.

Cost Function	S_1^η	S_1^{max}	S_2	S_{HMM}	S_{GMM}
Cnorm$_A$	**0.130**	0.171	0.132	0.336	1
Cnorm$_B$	**0.245**	0.313	**0.245**	0.474	1

Table 4 presents the performance of the five scoring methods on the same experiments as before in terms of Equal Error Rates (EER). Performance are given for each of the three classes of imposture involved in text-dependent speaker verification separately. By looking at each class of imposture, we aim to better understand the effect of each modelling of the alternative hypothesis.

Table 4: Performances of the different scoring methods in terms of Equal Error Rate (%) when testing the target speaker pronouncing the correct pass-phrase against the three classes of impostor trials defined for text-dependent speaker verification.

Imposture definition	S_1^η	S_1^{max}	S_2	S_{HMM}	S_{GMM}
$(\mathcal{X}, \overline{\mathcal{P}})$	1.51	**0.46**	1.68	4.57	50
$(\overline{\mathcal{X}}, \mathcal{P})$	1.75	2.22	1.75	**1.60**	4.92
$(\overline{\mathcal{X}}, \overline{\mathcal{P}})$	0.24	**0.20**	0.25	0.37	5.04

Results from the last column of Table 4 illustrate the fact that S_{GMM} score is not designed for text-dependent speaker verification. Thus, it does not consider the pass-phrase information and get an EER of 50% for the case of target speaker pronouncing a wrong pass-phrase, $(\mathcal{X}, \overline{\mathcal{P}})$, as it cannot separate the target speaker pronouncing the correct pass-phrase from the target speaker pronouncing a wrong pass-phrase. This phenomenon explains the high minimum cost values observed previously for S_{GMM}.

S_1^η and S_2 which were shown to minimize the cost functions, do not minimize EER for any of the imposture class. However, they don't either maximize EER in any condition but rather offer a good compromise for all classes of impostures. This compromise is reflected by the minimization of the cost functions.

One interesting observation is that the lowest EER when dealing with target speaker pronouncing a wrong pass-phrase (row 1 of Table 4) belongs to S_1^{max} score. For each trial, this scoring method selects the more likely imposture class to model the alternative hypothesis. A deeper analysis shows that for 99.07% of the TAR-wrong trials, S_1^{max} models the alternative hypothesis by using the speaker-dependent text-independent model λ_{gmm}. This correct detection of the class of imposture allows the S_1^{max} score to reduce the EER by 90% relatively to the S_{HMM} score for this class of impostures

Performance against impostor pronouncing a wrong pass-phrase $(\overline{\mathcal{X}}, \overline{\mathcal{P}})$ are given in the third row of Table 4 for reference. These results are consistent with the previous conclusions. As expected, the equal error rate is low for this condition as neither the identity nor the lexical content is correct.

5. DISCUSSION

In this paper, we have described text-dependent speaker verification as a classification task involving four classes of trials. We proposed to compute the verification score as a log-likelihood ratio in which the two competing components are defined by considering those four classes of trials. More specifically, we proposed three formulations of the alternative hypothesis score, written as a combination of the scores computed for the different classes of impostures separately.

The three scores resulting from our work have been compared to two existing scores: the classical GMM/UBM text-independent score and its text-dependent equivalent proposed for the HiLAM architecture in [18, 11]. Experiments conducted on the RSR2015 database have shown that the proposed scores outperform the two baseline scores. Indeed, the best scoring method proposed in this work decreases the two minimum costs considered by 61% and 48% relatively to our text-dependent baseline.

Interestingly, we observed that scoring methods minimizing the cost functions do not obtain the lowest equal error rate for any of the imposture class. Future work will focus on the different characteristics of the proposed scoring methods in order to minimize the error rate for each class of imposture separately.

6. REFERENCES

[1] Helen M. Wood, "The use of passwords for controlling access to remote computer systems and services," in *National Computer Conference*. 1977, pp. 27–33, ACM.

[2] Matthieu Hébert, *Text-dependent speaker recognition*, Springer-Verlag, Heidelberg, 2008.

[3] Anthony Larcher, Pierre-Michel Bousquet, Kong Aik Lee, Driss Matrouf, Haizhou Li, and Jean-Francois Bonastre, "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2012, pp. 4773–4776.

[4] Hagai Aronowitz, "Text-Dependent Speaker Verification Using a Small Development Set," in *Odyssey Speaker and Language Recognition Workshop*, 2012.

[5] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[6] Frederic Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meigner, Teva Merlin, Javier Ortega-Garcia, Dijana Petrovska-Delacretaz, and Douglas A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, April 2004.

[7] Douglas A. Reynolds and Larry Heck, "Integration of Speaker and Speech Recognition Systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1991, pp. 869–872.

[8] Larry Heck and Dominique Genoud, "Integrating Speaker and Speech Recognizers: Automatic Identity Claim Capture for Speaker Verification," in *Odyssey Speaker and Language Recognition Workshop*, 2001, pp. 249–254.

[9] Achintya Kumar Sarkar and S. Umesh, "Investigation of Speaker-Clustered UBMs based on Vocal Tract Lengths and MLLR matrices for Speaker Verification," in *Odyssey Speaker and Language Recognition Workshop*, 2010, p. 13.

[10] Wei-Qiang Zhang, Yuxiang Shan, and Jia Liu, "Multiple Background Models for Speaker Verification," in *Odyssey Speaker and Language Recognition Workshop*, 2010, p. 9.

[11] Kong Aik Lee, Anthony Larcher, Helen Thai, Bin Ma, and Haizhou Li, "Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 3317–3318.

[12] Shigeru Katagiri, Biing-Hwang Juang, and Chin-Hui Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2345–2373, 1998.

[13] Jinyu Li, Sibel Yaman, Chin-Hui Lee, Bin ma, Rong Tong, Donglai Zhu, and Haizhou Li, "Language Recognition Based on Score Discrimination Feature Vectors and Discriminative Classifier Fusion," in *Odyssey Speaker and Language Recognition Workshop*, 2006.

[14] Chin-Hui Lee, *Principles of Spoken Language Recognition*, Springer Handbook of Speech Processing, 2008.

[15] Niko Brümmer, Lukas Burget, Jan Honza Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karafiat, David A. Van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[16] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[17] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1994, vol. 2, pp. 291–298.

[18] Anthony Larcher, Jean-Francois Bonastre, and John 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.

[19] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012, pp. 1580–1583.

[20] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent Speaker Verification: Classifiers, Databases and RSR2015," *Speech Communication*, 2014, Accepted for publication.

[21] National Institute of Standards and Technology, "The nist year 2012 speaker recognition evaluation plan," `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`, 2012.