A NEW STUDY OF GMM-SVM SYSTEM FOR TEXT-DEPENDENT SPEAKER RECOGNITION

Hanwu SUN, Kong Aik LEE and Bin MA

Human Language Technology Department, Institute for Infocomm Research,
A*STAR, Singapore 138632

{hwsun, kalee, mabin}@i2r.a-star.edu.sg

ABSTRACT

This paper presents a new approach and the study of GMM-SVM system for text-dependent speaker recognition on scenario of the fixed pass-phrases. The uniform-split content-based GMM-SVM system is proposed and applied to text-dependent speaker evaluation. We conducted detailed study of the proposed method compared to the baseline GMM-SVM system on the RSR2015 database, which has been designed and collected for the evaluation of text-dependent speaker verification system. The experiment results show that the new approach can significantly reduce the detection error of the target-wrong error type (i.e., target speaker with wrong pass-phrase) while maintaining a low detection error for both imposter-correct and imposter-wrong error types (i.e., imposter with correct pass-phrase and imposter with wrong pass-phrase). We also show that score normalization could be applied with respect to the imposter-wrong distribution as opposed to the imposter-correct distribution.

Index Terms: speaker recognition, text dependent, channel compensation.

1. INTRODUCTION

With the increasing use of the mobile device and smart phones, user authentication based on short-utterance is required. Voice biometrics provides a potential method to authenticate whether the speaker is true speaker. The text-dependent speaker recognition is widely used for it [1, 2, 3, 4, 5]. Unlike text-independent speaker verification system [6], which is a process of verifying the identity without constraint on the speech content, text-dependent speaker verification requires the speaker pronouncing pre-determined pass-phrase [1, 2, 3, 4]. These pass-phrases may be unique, user dependent, or prompted by the system [1, 3, 4], as detailed below:

UNIQUE PASS-PHRASE: each client utters the same pass-phrase. This is the most constrained scenario as both duration and text are fixed [1, 4].

USER-DEPENDENT PASS-PHRASE: each client pronounces his own pass-phrase (chosen or generated by the system). In this scenario, duration and lexical content vary between speakers [1, 4].

PROMPTED TEXT: each client pronounces a sentence prompted by the system. This case does not require the user to remember a specific pass-phrase and reduces the risk of replay attacks. Duration variability can be easily reduced by adding a constraint on the prompts while lexical variability can be decreased by limiting the phonetic content of the prompts. A very common approach consists of using series of randomly ordered digits [1, 4].

In recent years, we have seen reviving interest on text-dependent speaker recognition [1, 2, 3, 5]. In [2], text-dependent speaker verification results were reported for the joint factor analysis (JFA), GMM-SVM-NAP, and GMM-HMM-NAP techniques based on a corpus collected by Wells Fargo Bank. In that paper, the GMM-HMM-NAP system was assisted by using the available utterance transcription. The Viterbi based forced alignment can also be applied [1, 3]. Obviously, the utterance transcription may not always be available for real-time application. Forced alignment via Viterbi decoding may also introduce some errors in the utterance segmentation. The condition will become worse if the utterance is corrupted by noise. In addition, both approaches introduce additional complexity to the speaker verification system.

In this paper, we introduce a simple uniform-split content-based GMM-SVM-NAP approach and experiment on the *Robust Speaker Recognition* 2015 (RSR2015) database [1, 3]. RSR2015 database was designed for the simulation and comparison of speaker verification systems in user-dependent English pass-phrase use-case. In addition, we also investigate the effect on different number of training utterances for the text-dependent system performance. The benefits of the score normalization and channel compensation on our proposed system are also investigated and discussed.

This paper is organized as follows: we first give an overview of the RSR2015 databases for text-dependent speaker verification in Section 2. Section 3 describes the GMM-SVM-NAP speaker recognition system for the text-dependent evaluation. The proposed content-based GMM-SVM speaker recognition system is presented in Section 4. Section 5 presents the detailed analysis of the GMM-SVM verification system on the RSR2015 database. Finally, we give the conclusions in Section 6.

Table 1: Different types of trials considered for experiments [1].

	Target speaker	Imposter speaker
Correct Pass-phrase	Target-Correct	Imposter-Correct
Wrong Pass-phrase	Target-Wrong	Imposter-Wrong.

2. THE RSR2015 DATABASE

The RSR2015 database [1, 3, 8] contains audio recordings from 300 persons, 143 female and 157 male speakers. Participants were selected to be representative of the ethnic distribution of Singaporean population. Selected speakers were between 17 to 42 years old. Each of the participants recorded nine sessions using three portable devices. Each session consists of thirty short sentences [1, 8].

The database was collected in office environments using six portable devices (four smart phones and two tablets) from different manufacturers. Each speaker was recorded using three different devices out of the six. The three devices were labelled as A, B and C. The following recording sequence was applied for each speaker: A, B, C, A, B, C, A, B, C, total 9 sessions. For each session, a speaker read thirty short sentences. These sentences were selected from TIMIT database [9] to cover all English phones. The number of words per sentence varies from four to eight. The average duration is 3.2 seconds, which include the beginning and ending silence sections.

Table 1 shows four types testing detection trials consider use case of user-dependent pass-phrases scenario. If we combine Target-Correct trials with other three imposter trials, we can have the following three subtasks: Imposter-Correct, Target-Wrong and Imposter-Wrong. For brevity, we denote these as IMP-CORR, TAR-WRG, and IMP-WRG, respectively.

3. GMM-SVM-NAP FOR TEXT-DEPENDENT SPEAKER RECOGNITION

The speaker recognition system used in this study is based on the GMM-based *support vector machine* (GMM-SVM) and the *nuisance attribute projection* (NAP) technique for channel compensation, in short, the GMM-SVM-NAP as reported in [7]. In this approach, speech utterances with variety of durations are represented as high-dimensional vectors referred to as the GMM supervectors. Channel compensation and speaker detection are then performed in the high-dimensional vector space.

Let $\Lambda = \{ \omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i; i = 1, 2, \cdots M \}$ be the parameters of the universal background model (UBM), where M is the number of mixture components, ω_i are the mixture weights, $\mathbf{m}_i$ are the mean vectors, and $\boldsymbol{\Sigma}_i$ are the covariance matrices assumed to be diagonal. For a given utterance X_s, the Baum-Welch statistics are used to adapt the mean vectors of the UBM using the maximum *a posteriori* (MAP) [12]. The adapted mean vectors are concatenated to form a GMM supervector.

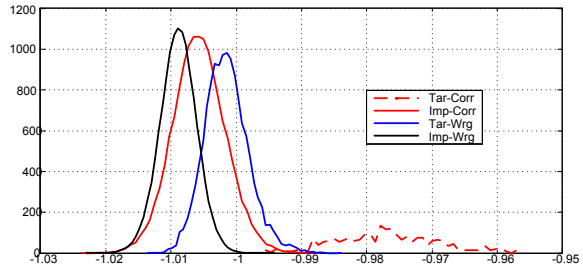

Figure 1: The score distribution of the GMM-SVM-NAP system used for text-dependent task exhibits four distinct modes.

Based on the NAP matrix derived from the training dataset, the supervectors are channel compensated via NAP projection. For the case of text-dependent speaker recognition, the NAP matrix has to be trained in a text-dependent manner, where speaker-sentence pair is taken into consideration. More specifically, utterances are grouped per speaker-sentence (i.e., multiple utterances of the same sentence by a speaker) in constructing the within-class covariance matrix. The NAP projection matrix is derived from eigenvalue decomposition of the covariance matrix, from which the eigenvectors (with largest eigenvalues) are concatenated to form the so-called NAP matrix.

Another point to note for the case of text-dependent speaker verification is that we have to deal with three types of non-target trials, as opposed to one for the case of text-independent. Figure 1 shows a typical score distribution obtained with a GMM-SVM-NAP system. The score distribution exhibits four distinct modes. This greatly affects the strategy that could be used for score normalization [13]. Recall that score normalization is performed with respect to imposter distribution. It is general difficult, if not impossible, to select cohort utterances that would satisfy all the three imposters distributions. One viable option is to base the score normalization on the IMP-WRG distribution, where cohort sentences could be selected from a background set of speakers with sentences different from that of the pass-phrases.

4. CONTENT-BASED GMM-SVM-NAP

In real-time applications, the transcription of the pass-phrases may not always be available. In addition, it will significantly increase the complexity of the speaker recognition system. Automatic forced alignment via Viterbi decoding will unavoidably produce errors. Such errors will significantly increase if the recorded pass-phrases have low signal-to-noise ratio. In text-dependent speaker verification system, the required system performances are usually very high. High error rate propagated from the upstream forced alignment is always unfavourable.

Contrary to the HMM-SVM system reported in [1, 2, 3], which requires the transcription for the given utterances or the forced alignment via Viterbi decoding to conduct experiment, we introduce a content based GMM-SVM-NAP. The content based GMM-SVM-NAP system is based on the idea of uniform-split pass-phrase without the knowledge of utterance transcription or forced alignment. We split an utterance into *N* equal segments and extract *N* supervectors from the utterance, one for from each segment. These supervectors are concatenated to form a single content-based supervector. This operation is applied to the model training, testing, SVM background dataset, and the t-norm dataset. These content-based supervectors are then taken as input to the GMM-SVM-NAP system. For example, *split1* is to uniform split the utterance feature into left and right two equal parts. Then, we concentrate these two supervectors into one double size supervector.

Let us see how a uniform-*split1* would behave on the four different types of trials as listed in Table 1. For the TAR-CORR trials, a speaker uttering pattern is usually similar for the training and testing. After applying uniform-*split1* (i.e., splitting an utterance into left and right context with equal length), its performance may be kept similar as the whole utterance situation. Even though the speaking rate might be faster or slower than his/her normal style, it may still get relatively consistent splitting at the middle alignment point. However, *split1* can significant affect the results of target speaker uttering wrong pass-phrase (i.e., the TAR-WRG trials). The TAR-WRG scores shown in Figure 1 are the closest to the target score distribution. Notice that such trials are target trials in the text-independent system, but they are imposter trials in text-dependent system. Since we reduce the training and test utterances into half length, such imposter scores will be of course reduced significantly.

Of course, it is obvious that there may be very big error for content matching or alignment if the high order uniform-split is applied. If we look at the three subtasks indicated in the text-dependent system in Figure 1, it is interesting to study how uniform-split helps these three individual subtasks.

5. SPEAKER RECOGNITION EXPERIMENT

We used RSR2015 database to conduct text-dependent speaker verification on GMM-SVM-NAP and the content based GMM-SVM-NAP system. The RSR2015 datasets are simple divided into disjoint partitions (in terms of speakers) of equal size disjoint. The first partition is used as the training/test to evaluate the text-dependent system performance. The second partition is used for the GMM model training, NAP dependent design, as well as SVM system background dataset.

We use the MFCC feature in this study. In particular, 19-dimenison MFCC features are generated for each speech frame with a window of 30ms and a frame shift of 12.5ms. By including the 19-dimension of the first derivatives and the 12-dimension of the second derivatives, a MFCC feature vector consists of 50 dimensional features. The spectral subtraction technique is used to assist the voice activity detection (VAD) for selecting useful speech frames [10]. The MFCC feature vectors are then processed by RASTA filtering [11] and followed by *mean and variance* normalization (MVN).

Table 2: Number of target and imposture trials for male and female subsets for each of 30 pass-phrases in Partition I.

	Target-correct	Imposter-correct	Target-wrong	Imposter-wrong
Male	474	36,972	13,746	1,072,188
Female	432	30,672	12,528	889,488

Based on the above dataset selection, Partition I consist of 79 male and 72 female speakers available for the test set. Each speaker has 30 different pass-phrases. Each pass-phrase has up to 3 sessions for training and 6 sessions for test. The numbers of target and non-target trials are shown in Table 2. For training a model, we use sessions 1, 4, and 7 for the speakers in Partition I. The remaining 6 sessions {2, 3, 5, 6, 8 9} are used for the test [1, 3]. So each speaker is enrolled with only one mobile device [1, 3]. As mentioned earlier in Section 2, the device used for sessions {1, 4, 7} is of the same type.

Based on the above setting, we conducted three experiments on RSR2015 database. The first experiment studies the effects of the number of training sessions on the performance of short pass-phrases text-dependent speaker verification. Secondly, we examined the proposed uniform-split GMM-SVM-NAP system in comparison to the whole utterance case GMM-SVM-NAP system. Finally, we further investigate any benefit of score normalization and channel compensation on short pass-phrase speaker verification system. We use the equal error rate (EER) [6], averaged over 30 sentences, to evaluate the system performance. In the experiments, the size of UBM is set to 256.

5.1 Effects of Number of Training Utterances for GMM-SVM System

It is commonly known from NIST SREs [6] that multiple-session enrollment leads to a much better performance than single-session training. It is not surprise that increasing the number of the training utterances improves the performance as well for text-dependent system. The results are shown Table 3 for the case of 1, 2 and 3 sessions of the same pass-phrase available for enrollment. From Table 3, it is evident that a significant improvement is obtained with 2 enrollment sessions as compared to 1 session. Increasing the number of enrollment sessions to 3 brings further improvement, however, the impact start to level off. Since the pass-phrase is usually short, it is very easy to obtain two or more pass-phrases for training a target speaker model for a given pass-phrase model.

Table 3: Performance of the proposed system under different number of training utterances in terms of absolute EER (%) and percentage of relative improvement (Impro).

	No of sessions	Male		Female	
		EER	Impro.	EEE	Impro.
IMP-CORR	1	2.219	-	1.388	-
	2	0.642	71%	0.297	78%
	3	0.464	79%	0.180	87%
TAR-WRG	1	3.102	-	1.520	-
	2	1.292	58%	0.436	71%
	3	0.977	69%	0.244	84%
IMP-WRG	1	0.680	-	0.368	-
	2	0.170	75%	0.043	88%
	3	0.137	80%	0.029	92%

Table 4: Performance of the proposed system for different number of splits *N* in terms of absolute EER (%) and the percentage of relative improvement (impro).

	N	Male		Female	
		EER	Impro.	EEE	Impro.
IMP-CORR	0	0.464	-	0.180	-
	1	0.455	2%	0.167	7%
	2	0.539	-16%	0.250	-38%
	3	0.553	-19%	0.281	-56%
TAR-WRG	0	0.977	-	0.244	-
	1	0.611	37%	0.145	40%
	2	0.562	42%	0.166	32%
	3	0.506	48%	0.198	19%
IMP-WRG	0	0.137	-	0.029	-
	1	0.125	8.7%	0.025	14%
	2	0.135	2%	0.035	-20%
	3	0.132	4%	0.059	-103%

5.1 Content-based SMM-SVM System Results

From the EERs shown in Table 3, we notice that the worst performance occurs for the TAR-WRG trials. The EERs are much higher than those of the IMP-CORR and IMP-WRG. In addition, the performance gap enlarges with more training utterances used. Recall that the TAR-WRG trials are considered as target trials in text-independent speaker verification.

To evaluate the benefit of the proposed content-based GMM-SVM-NAP system, we use the whole utterance GMM-SVM-NAP system as baseline. We split the utterance up to 4 uniform splits ($N = 3$). Here, three training sessions were used for the enrollment. The t-norm and NAP were applied. Table 4 and Figure 2 showed the uniform-split GMM-SVM system performances. From the results shown in Table 4, *split1* improves the performance across all three subtasks compared to our baseline whole utterance condition. More importantly, the EER for the TAR-WRG subtask was significantly reduced by 37% and 40% for male and female, respectively. We also notice that uniform splitting does not work well for 3 and 4 splits, especially on IMP-CORR. Overall, the best performance setting is the uniform *split1* where utterances were split into left and right context segments.

5.3. NAP, Score Normalization Results

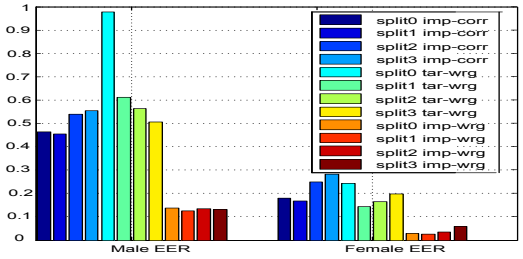

Figure 2: Male and female averaged EER by applying different uniform-split supervectors.

Table 5: Averaged EER by applying different channel and score normalization.

		Male		Female	
		EER	Impro.	EEE	Impro.
IMP-CORR	raw	0.639	-	0.277	-
	t-norm	0.583	8.7%	0.251	9%
	NAP	0.475	26%	0.202	27%
	NAP + t-norm	0.455	29%	0.167	40%
TAR-WRG	raw	0.984	-	0.333	-
	t-norm	0.892	9%	0.224	33%
	NAP	0.817	17%	0.212	36%
	NAP + t-norm	0.611	38%	0.145	56%
IMP-WRG	raw	0.155	-	0.028	-
	t-norm	0.132	15%	0.025	11%
	NAP	0.130	16%	0.022	21%
	NAP + t-norm	0.125	19%	0.019	32%

We further studied the effect of score normalization and channel compensation on the uniform-*split1* content-based GMM-SVM-NAP system. The raw scores were as our baseline. The results with the t-norm, NAP, and NAP followed by t-norm were evaluated as shown in Table 5. It can be seen that the t-norm alone achieved 8% to 33% of relative improvement in the three subtasks. In addition, the NAP, which is extracted from text-dependent utterances, can significantly improve all three subtasks, especially for IMP-CORR and TAR-WRG task from 16% to 36% improvement. We also observe that the best performance of all the three subtasks is achieved by applying both NAP and t-norm, with maximum up to 56% EER improvement.

6. CONCLUSIONS

We have presented a detailed study for GMM-SVM-NAP system on the RSR2015 database. Without using the prior knowledge of utterance transcription and forced alignment, the proposed uniform-split context-based approach can significantly reduce the TAR-WRG subtask, while keeping or slight improving the IMP-CORR and IMP-WRG subtasks. The experiment results also confirmed that text-dependent NAP compensation can obviously improve all the three subtasks in EER of system performances. Meanwhile, it is advisable to use more than one enrolment sessions to train a given pass-phrase speaker model to achieve the reasonable text dependent speaker recognition performance for short utterance biometric applications.

7. REFERENCES

[1] A. Larcher, K. Lee, B. Ma, and H. Li, "Text dependent speaker verification: classifiers, databases, and RSR2015", Speech Communication, 60, pp. 56-77, Apr. 2014.

[2] H. Aronowitz, R. Hoory, J.Pelecanos and D. Nahamoo, "New Developments in Voice Biometrics for User Authentication". in *Proc. INTERSPEECH*, pp. 28-31, Florence 2011.

[3] A. Larcher, K. Lee, B. Ma, and Haizhou Li, "RSR2015: database for text-dependent speaker verification using multiple pass-phrases", in *Proc. INTERSPEECH*, , Sep. 2012.

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

[5] M. F. BenZeghiba and H. Bourlard, "User-customized password speaker verification using multiple reference and background models," Speech Communication, vol. 48, no. 9, pp. 1200–1213, 2006.

[6] NIST 2008 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf.

[7] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, pp. 97–100, 2006.

[8] "RSR2015 Overview & Specifications", https://www.etpl.sg/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus linguistic data consortium," Philadelphia, PA, vol. 1, 1993.

[10] H. Sun, B. Ma and H. Li, "An Efficient Feature Selection Method for Speaker Recognition," in *Proc. ISCSLP*, pp. 181–184, 2008.

[11] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[12] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, 10(1):19-41, 2000.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no 1-3, pp. 42–54, Jan 2000.