CSLT TECHNICAL REPORT-20150017 [Thursday 15th October, 2015]

AN OPEN/FREE DATABASE AND BENCHMARK FOR UYGHUR SPEAKER RECOGNITION

Askar Rozi[??,??]
, Dong Wang[??*]
and Zhiyong Zhang[??]

[*]Correspondence: wang-dong99@mails.tsinghua.edu.cn
[??]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Few research has been conducted on Uyghur speaker recognition. Among the limited works, researchers usually collect small speech databases and publish results based on their own private data. This 'close-door evaluation' makes most of the publications doubtable. This paper publishes an open and free speech database THUYG-20 SRE and a benchmark for Uyghur speaker recognition. The database is based on the THUYG-20 speech corpus we recently released, and the benchmark involves recognition tasks with various training/enrollment/test conditions. We provide a complete description for the database as well as the benchmark, and present an i-vector baseline system constructed using the Kaldi toolkit.

Keywords: Uyghur; THUYG-20; speaker recognition

1 Introduction

Speaker recognition (SR) authenticates the claimed identity of a person by speech input. The GMM-UBM approach was the dominant technology in 90's, and tody's state-of-the-art is the i-vector approach. The US National Institute of Standards and Technology (NIST) has organized a series of Speaker Recognition Evaluations (SRE) with standard databases and evaluation protocols. These evaluations provide a standard benchmark for researchers to evaluate their work and compare with each other. It significantly promotes the development of speaker recognition technologies. After a decade of research, current speaker recognition systems have attained rather satisfactory performance [**?**, **?**].

Despite the excellent improvement achieved in NIST SRE, few research works have been conducted in the field of Uyghur speaker recognition. Among the limited research, most of thew work focus on small modifications of the GMM-UBM framework that has been out of date. For example in [**?**], a modified vector quantisation method was proposed instead of the conventional GMM-UBM. In [**?**], a GMM-UBM/SVM approach was proposed to leverage the robustness of GMM/UBM against noise and the discriminative nature of SVM in scoring. The same method was also described in [**?**]. To the best of our knowledge, current state-of-the-art speaker recognition technologies such as JFA and i-vectors have not been studied in Uyghur speaker recognition.

More seriously, these limited works on Uyghur SR are based on small databases that are collected and used by individual researchers 'privately'. For example, the database collected by Reyiman et al. consists of 350 speakers [**?**], and in [**?**], the experiments were conducted with 70 target speakers. Li et al. conducted their research on a database consisting of 50 speakers [**?**]. These databases are not publicly available, which makes the publications not reproducible by others and compared with each other. A standard speech database that is open and free is highly desirable.

In a previous study, we published a free speech database THUYG-20 for Uyghur speech recognition [**?**]. This database consists of 348 native Uyghur speakers and can be used for speaker recognition. In this paper, we publish THUYG-20 SER, a database based on THUYG-20 but re-designed specifically for speaker recognition. Based on this database, we setup a benchmark for speaker recognition which involves a set of SR tasks in various training/enrollment/test conditions. Additionally, a baseline system based on the modern i-vector technology is constructed and the recipe and results are published online. We provide the complete data description, system architecture, experimental set up and evaluation performance. These can be used as a full reference for Uyghur speaker recognition research. The database can be downloaded from http://data.cslt.org/thuyg20-sre/README.html.

The rest of the paper is organized as follows: Section 2 brief introduces the i-vector technology, and Section 3 presents the THUYG-20 SRE database and the benchmark. The baseline system is presented in Section 4, followed by some conclusions in Section 5.

2 I-vector technology

2.1 I-vector

Given an utterance, the i-vector model assumes that the speaker-dependent supervector M is 'generated' by:

$$M = m + Tw \tag{1}$$

where m is a speaker and channel independent supervector, T is a low-rank matrix, and w is a low-dimensional vector that represents the utterance. Assuming that w follows a standard normal distribution $N(0, I)$, Eq. (1) is a linear Gaussian model and M follows a Gaussian distribution $N(m, TT^T)$. The parameter estimation and variable inference with this model can be easily performed. Specifically, given a set of training speech signals $\{X_i\}$, the matrix T is estimated by optimizing the following likelihood function:

$$\mathcal{L}(T) = \sum_i ln\{P(X_i; T)\} = \sum_i ln\{\sum_M P(X_i; M)P(M; T)\}$$

where the conditional probability $P(X_i; M)$ is modeled by a GMM, and the prior probability $P(M; T)$ is a Gaussian. Once T is estimated, inferring the posterior probability of w given an utterance X is simple since $P(w|X)$ is a Gaussian as well. In most cases only the mean vector (so called 'i-vector') of the posterior is

concerned and it can be obtained by maximum a posterior (MAP). More details of the computation can be found in [**?**].

With utterances represented by i-vectors, the score of a test speech on a claimed speaker can be derived as the cosine distance between the i-vectors of the test speech and the enrollment speech of the claimed speaker.

2.2 Probabilistic LDA

The i-vector model is a total-variability model which means that an i-vectors represents both speaker characteristics and other non-speaker factors particularly channels. This is certainly unideal for discriminating speakers. Probabilistic LDA (PLDA) separates the total-variability space into a speaker subspace and a channel subspace, so that speakers can be represented more accurately. This model can be formulated by:

$$w_r = m + Ux_r + Vy + \epsilon_r \tag{2}$$

where w_r is the i-vector of the r-th utterance, and m is the population mean. U represents the channel subspace and x_r is a channel vector; V represents the speaker subspace and y is a speaker vector. Finally, ϵ_r represents the residual. Note that x_r and y follow the standard Gaussian distribution, and ϵ_r follows a Gaussian distribution $N(0, \Sigma)$. The parameters $\{m, U, V, \Sigma\}$ can be estimated using the EM algorithm, and the inference for the speaker vector y can be achieved by MAP. Scoring a test speech can be performed with the speaker vectors using cosine distance, though a full Bayesian approach is more often used [**?**].

3 THUYG-20 SRE database and benchmark

3.1 THUYG-20 SRE database

Recently, we published an open Uyghur speech database THUYG-20 which is totally free for researchers.[1] This database consists of more than 20 hours of speech signals spoken by 371 speakers. The signals were recorded in a silent office by the same carbon Microphone. The sample rate is 16000 Hz and the sample size is 16 bits. The speakers were all colleague students at the age of 19-28, and they are native Uyghur speakers from 30 counties. The sentences were excerpted from general domains including novels, newspapers and various types of books. The database were recorded from January to September in 2012. More details can be found in [**?**].

Although the original purpose of THUYG-20 was for speech recognition, it can be used for speaker recognition as well. We publish the THUYG-20 SRE database that is based on THUYG-20 but re-designed for speaker recognition. The data profile of the database is shown in Table 1. The entire database is split into three datasets: the training set involves 4771 utterances spoken by 200 speakers, and is used to train models including the UBM in the GMM-UBM framework, the the T matrix in the i-vector model, and the parameters $\{m, U, V, \Sigma\}$ in PLDA. The enrollment and test sets consist of the same set of 153 speakers. Each enrollment utterance is 30 seconds and each test utterance is 10 seconds.

[1] http://data.cslt.org/thuyg20/README.html

Dataset	Spk.	Female	Male	Utt.	Dur. (hrs)
Training	200	100	100	4771	13.15
Enrolment	153	87	66	153	1.28
Test	153	87	66	2361	6.56

Table 1: The data profile of THUYG-20 SRE. 'Spk.' denotes the number of speakers, 'Utt.' denotes the number of utterances, 'Dur.' denotes the duration of the speech signals in hours.

Additionally, the database involves three noise signals obtained from the DEMAND noise database[2]: white noise, cafeteria noise and car noise. A script is provided to mix the noise to the speech signals in a random fashion, with a specified SNR level.

3.2 SRE benchmark

Based on THUYG-20 SRE, we propose a benchmark for Uyghur SRE as follows:

- The evaluation is categorized into two classes: the limited-resource test and the open-resource test. In the limited-resource test, all the models are trained with THUYG-20 SRE only, and in the open-resource test, any data are allowed to train the models.
- The evaluation is gender-dependent, following the convention of the NIST SRE evaluation plan [**?**].
- The evaluation is conducted under three noise types: white noise, cafeteria noise and car noise. The enrollment data and the test data can be corrupted by the same noise type only, however the corruption can be in different SNRs, selected from (-6, -3, 0, 3, 6, 9, clean), where 'clean' means no noise corruption.
- The evaluation considers three enrollment conditions for which the length of the enrollment speech is 10, 20 and 30 seconds respectively. The required length of enrollment speech is obtained by cutting the enrollment utterance from the beginning. Note that the test speech is fixed to be 10 seconds.

As a quick summary, the THUYG-20 SRE evaluation is a set of speaker recognition tasks, each of which is specified by the data resource, the gender, the noise type, the enrollment SNR, the test SNR, and the length of the enrollment speech. It is a large set but interested researchers can select part of them in their study. A major contribution of this paper is to present the baseline results for these tasks that researchers can compare to and compete with. We also call challenge on these tasks and record the state-of-the-art results on the challenge web site.[3]

4 Baseline system results

This section describes our baseline system and reports the performance it achieved on the THUYG-20 SRE tasks. Due to the large number of tasks, we just report the results with female speakers and with cafeteria noise. The full set of results can be found in the challenge web site.

[2]http://parole.loria.fr/DEMAND/
[3]http://cslt.riit.tsinghua.edu.cn:8081/data/thuyg20-sre/challenges/sre.html

		EER%		
Enr. SNR	Test SNR	C10	C20	C30
clean	clean	6.35	5.11	4.01
clean	9db	16.56	17.43	17.43
clean	0db	26.26	27.64	27.43
9db	clean	15.68	11.96	10.94
9db	9db	12.25	9.92	8.83
9db	0db	17.80	16.19	15.75
0db	clean	26.70	19.91	18.82
0db	9db	17.07	14.08	13.57
0db	0db	18.16	17.94	17.58

Table 2: EER results of the i-vector + PLDA baseline where models are trained with THUYG-20 SRE. 'Enr.' stands for 'Enrollment'. 'C10', 'C20' and 'C30' denote three enrollment conditions where the enrollment speech is in 10, 20 and 30 seconds respectively.

4.1 System configuration

The baseline speaker recognition system we built is based on the state-of-the-art i-vector framework, which involves the i-vector model for speaker vector extraction and the PLDA model for channel compensation. The Mel frequency cepstral coefficients (MFCCs) are used as the features, which involves 20-dimensional static MFCCs and the Δ and $\Delta\Delta$ dynamic coefficients, resulting in MFCC vectors of 60 dimensions. To remove channel effect, cepstral mean and variance normalization (CMVN) has been applied at the utterance level. The UBM involves 2048 Gaussian components, and the i-vector dimension is set to 400. The experiments were performed using the Kaldi toolkit [**?**].

4.2 Limited-resource task

The first experiment examines the performance of the baseline system on the limited-resource task, i.e., only the data in THUYG-20 SRE are used for model training. The results on female speakers are reported in Table 2, where the corruption is cafeteria noise, and only three SNR conditions are reported (clean, 9db and 0db). As shown in Table 1, there are 87 speakers in the test; these speakers are tested against each other, resulting in $119,277$ trials in total.

From the results in Table 2, it can be seen that with clean enrollment and test speech, the performance is rather good, even the enrollment speech is as short as 10 seconds. With noise corruption, the EERs are significantly increased, no matter if the corruption is on enrollment or test speech. More heavy the corruption is, more significant the performance reduction is observed. If the enrollment utterance is relative long (i.e., 30 seconds), the impact of noise corruption on test speech is more evident than on enrollment speech. For example, in the case of SNR=9db, the EER is 10.94 with the corruption on enrollment data, while the number is 17.43 with the corruption on test data. Additionally, if the SNR level matches for enrollment and test, the performance tends to be less impacted.

To improve the performance in noisy conditions, one can involve the same corruption in model training. This 'noisy training' can significantly improve performance in conditions where the enrollment and/or test speech are corrupted by the same level of corruption. Table 3 presents the performance with training data corrupted by cafeteria noise at SNR=9db. It can be seen that compared to the results in Table 2, the performance in noisy conditions is generally improved, particularly if the

		EER%		
Enr. SNR	Test SNR	C10	C20	C30
clean	clean	8.90	7.00	5.40
clean	9db	12.33	9.63	7.88
clean	0db	14.73	13.42	12.69
9db	clean	10.36	9.04	8.46
9db	9db	7.73	6.78	5.69
9db	0db	9.48	9.04	8.17
0db	clean	13.79	10.21	9.41
0db	9db	8.97	6.57	5.54
0db	0db	9.77	7.80	7.22

Table 3: EER results of the i-vector + PLDA baseline where models are trained with THUYG-20 SRE. The training data are corrupted by cafeteria noise at SNR=9db. The notations are the same as in Table 2.

		EER%		
Enr. SNR	Test SNR	C10	C20	C30
clean	clean	5.03	2.92	2.33
clean	9db	8.24	5.62	4.45
clean	0db	13.86	10.21	8.83
9db	clean	8.61	5.62	4.01
9db	9db	6.71	4.67	3.72
9db	0db	9.26	6.49	5.76
0db	clean	13.64	9.12	6.57
0db	9db	9.12	5.98	4.74
0db	0db	9.70	6.42	5.47

Table 4: EER results of the i-vector + PLDA baseline where models are trained with Fisher. The notations are the same as in Table 2.

enrollment and/or test speech are corrupted at the same SNR level (9db) as the training data.

4.3 Open-resource task

The THUYG-20 SRE database is relatively small for UBM/i-vector/PLDA training. In the open-resource task, extra data are allowed to improve the model quality. In this study, the Fisher corpus (219.59 hours,female part) is used as the extra data to train the models. Note that Fisher is an English corpus, however it turns out that the models trained with it work pretty well for Uyghur speaker recognition, as shown in Table 4. This is on one hand demonstrates that SRE is largely language independent, and on the other hand indicates that a large training database is important for building speaker recognition systems.

5 conclusions

This paper published an open and free speech database THUYG-20 SRE that is used for Uyghur speaker recognition. Additionally, we published the THUYG-20 SRE benchmark based on this database, and presented the baseline results based on the state-of-the-art i-vector framework. We hope these publications can promote the speaker recognition research in Uyghur.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No.61271389 and NO.61371136 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302. Thanks also for the data preparation by Prof. Askar Hamdulla at Xinjiang University.