Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

DeepMine Speech Processing Database: Text-Dependent and Independent Speaker Verification and Speech Recognition in Persian and English

Hossein Zeinali [1,2], Hossein Sameti [2], Themos Stafylakis [3]

[1] Sharif DeepMine Ltd., Tehran, Iran
[2] Sharif University of Technology, Tehran, Iran
[3] Computer Vision Lab, University of Nottingham & Omilia Conversational Intelligence, UK
hsn.zeinali@gmail.com, sameti@sharif.edu, themos.stafylakis@nottingham.ac.uk

Abstract

In this paper, we introduce a new database for text-dependent, text-prompted and text-independent speaker recognition, as well as for speech recognition. DeepMine is a large-scale database in Persian and English, with its current version containing more than 1300 speakers and 360 thousand recordings overall. DeepMine has several appealing characteristics which make it unique of its kind. First of all, it is the first large-scale speaker recognition database in Persian, enabling the development of voice biometrics applications in the native language of about 110 million people. Second, it is the largest text-dependent and text-prompted speaker recognition database in English, facilitating research on deep learning and other data demanding approaches. Third, its unique combination of Persian and English makes it suitable for exploring domain adaptation and transfer learning approaches, which constitute some of the emerging tasks in speech and speaker recognition. Finally, the extensive annotation with respect to age, gender, province, and educational level, combined with the inherent variability of the Persian language in terms of different accents are ideal for exploring the use of attribute information in utterance and speaker modeling.

The presentation of the database is accompanied with several experiments using state-of-the-art algorithms. More specifically, we conduct experiments using HMM-based i-vectors, and we reaffirm their effectiveness in text-dependent speaker recognition. Furthermore, we conduct speech recognition experiments using the annotated text-independent part of the database for training and testing, and we demonstrate that the database can also serve for training robust speech recognition models in Persian.

1. Introduction

Robustness in text-independent speaker recognition depends heavily on the availability of large amounts of in-domain training data. Some of its most frequently used approaches, such as joint-factor analysis and i-vectors, owe their success not merely on their innovative probabilistic framework, but on the way they leverage the abundant data of NIST to learn subspaces over high-dimensional utterance representations, [1, 2]. In recent years, the advent of deep learning methods and the introduction of end-to-end architectures made the requirement of large amounts of in-domain data way more essential, [3, 4, 5]. Training such architectures is currently infeasible without tonnes of in-domain data, together with data-augmentation techniques.

In text-dependent speaker recognition, experiments with end-to-end architectures conducted on large proprietary databases have demonstrated their superiority over traditional approaches, [6]. Yet, contrary to text-independent speaker recognition, where the NIST repository and other recently introduced databases (e.g. [7]) in general suffice for training deep learning models in English, text-dependent speaker recognition lacks of large-scale publicly available databases. Several efforts have been made to collect data for text-dependent speaker verification, with RSR2015 [8] and RedDots [9] being the most well-known. In the former, speech data was collected from 300 individuals in a controlled manner, while in the latter, data collection was implemented via an Android application in a crowdsourcing scenario. Despite the efforts, neither of these projects succeeded in collecting speech material that could enable robust training of large-scale speaker recognition models.

Another area of speech processing that requires large amounts of data, as well as high variability with respect to speakers and dialects is automatic speech recognition (ASR). There are several databases for training ASR in certain languages (e.g. English, Mandarin, a.o.) and modern deep learning approaches are heavily dependent on them, [10]. However, there is still lack of annotated datasets in many other languages, one of which is Persian. Such a deficiency has a strong negative impact on the ecosystem of these countries, considering the vast number of services and applications built on top of ASR systems in western and other developed countries.

Based on these observations and inspired by the RedDots project, we decided to create a multi-purpose public speech database including a large number of speakers. The main goal of the project is to collect speech from at least a few thousand speakers, enabling research and development of deep learning methods. Speech data is collected from the participants online with user interfaces such as web or mobile applications (i.e. crowdsourcing). The DeepMine project started at the beginning of 2017, and after the designing of the database and the development of an Android application and server programs, data collection was begun at mid of 2017. The duration of the project was chosen to be one year but it is possible that it will be extended for several months. So far, more than 1300 speakers have participated in the project.

The rest of the paper is organized as follows. In Section 2, we describe the way data collection is organized and how jobs are distributed between server and client sides. In Section 3, we analyze the DeepMine database in terms of its composing parts, we discuss the relevant speech and speaker recognition tasks associated with each part and we provide the statistics of the current state of the database. Finally, in Section 4, we present some preliminary experiments on speech and speaker recognition, using state-of-the-art algorithms.

10.21437/Odyssey.2018-54

2. Data Collection Scenario

DeepMine is a data collection project based on crowdsourcing, and as such is has certain similarities with RedDots. However, there are several differences between the two projects. For data collection, an Android application was designed, parts of which were inspired by the Android application proposed in [11, 12]. To communicate with the Android recording application from the server side, a TCP server was designed from scratch in Java. The whole pipeline of the data collection project is shown in Figure 1. More specifically:

1. **Installing application and respondent registration:** The first step for contributing to the DeepMine project is to install the Android application. The participant fills a registration form in the application, his/her information is sent to the server and his/her identity number is received from it. A variety of information is asked from the respondents:

   - **Personal and contact information:** Useful for future communication with participants and winners announcement. This information will not be included in the database and participants are informed about this during the registration.
   - **Age and gender:** Useful for research on age estimation and gender identification, as well as for gender dependent data processing.
   - **Educational and English levels:** The educational level can be used to investigate its relation with performance on several tasks, e.g. speech recognition. The English level is used to determine whether or not a respondent can participate on the English parts, and it can also be used to examine the effect of English level on the performance of text-dependent speaker verification systems. Four levels are considered: No ability to read English, Elementary: able to read English, Intermediate: able to comprehend English, and finally Excellent: able to speak English fluently.
   - **Province:** This type of information enables us to determine the number of participants across different parts of Iran in the project.
   - **Accent:** One of the main advantages of this database is the coverage of different Iranian accents. Accent information can be used in speaker and speech recognition as side information, as well as in accent recognition.

   There is a ***Terms and Conditions*** agreement in the registration form which respondents have to accept. In this terms, it is mentioned that the data of the participants will be included in a public database in a totally anonymous fashion. Informing the participants about the use of their speech data and protecting their personal informations are of major concern.
2. **Getting sessions content:** After completing respondent's registration, the application sends the content of 8 sessions, each of which has exactly 24 phrases, based on the respondent's English level. The respondent can record the 8 sessions offline and send them to the server, once he/she is connected to the Internet. After completing the 8 sessions, 8 new ones are sent to the respondent by the server. Each respondent can contribute up to 64 sessions.
3. **Recording sessions by respondent:** After receiving the session contents, the respondent can perform audio recording at any time and everywhere, provided that there is a minimum of 8 hours between two consecutive sessions. After initiating the recording process, the respondent provides additional information about the ongoing recording, such as environment condition (quiet or noisy) as well as the condition of his/her voice (can be "normal", "just woke up" or "having temporary laryngeal problems" such as colds, a.o.). The beginning and the end of the recording are determined by the user, and he/she can listen to the recorded material and record it again if needed. After applying ASR on the recorded phrase to verify its content, the user is directed to the next phrase.
4. **Sending recorded utterances to the server and doing server-side processing:** After connecting to the Internet, the completed utterances are sent to the server and saved in a hierarchical structure. Finally, further processing is performed to verifying again their content.

It is worth noting that similarly to the RedDots project, the participants can record their utterances in any environment they choose. Therefore, realistic conditions for practical applications are well considered in this database.

Similarly to the RedDots project, we encourage the participants to use more than one device for different utterances. Although this is not feasible for most of the participants, the variability in the environment across recording sessions yields rich within speaker variability.

2.1. Client-Side Processing

In order to prevent invalid recordings, several online processing (i.e. during session recording) are performed on the client-side:

1. **Voice activity detection (VAD):** The first stage of processing is VAD, applied using a simple and fast energy-based method. If an utterance is classified as silence, a warning message is shown to the respondent, indicating that the current phrase should be recorded again.
2. **Checking the noise level:** After determining the silence parts, the signal-to-noise ratio is estimated and if it is below 5db, the application shows another warning message to the respondent to record the phrase again in a less noisy environment.
3. **Checking the minimum and maximum duration of an utterance:** The expected duration of a phrase is estimated based on the phonemes it contains and the average phoneme duration. If the duration of a recorded utterance is shorter than a specific threshold (based on the phrase average), a warning message is displayed to the respondent to record the phrase again.

If a recording passes the aforementioned checking, the next phrase is shown to him/her. If not, the respondent should record again the current phrase.

2.2. Server-Side Post-processing

The following processing will be performed on the final version of the database before its release:

1. All client-side processing will be repeated with better and more accurate methods.
2. Utterance verification will be performed on text-dependent parts and problematic utterances will be detected.
3. Isolated word recognition will be performed on the text-prompted parts to detect mispronunciations or to remove incorrect utterances.

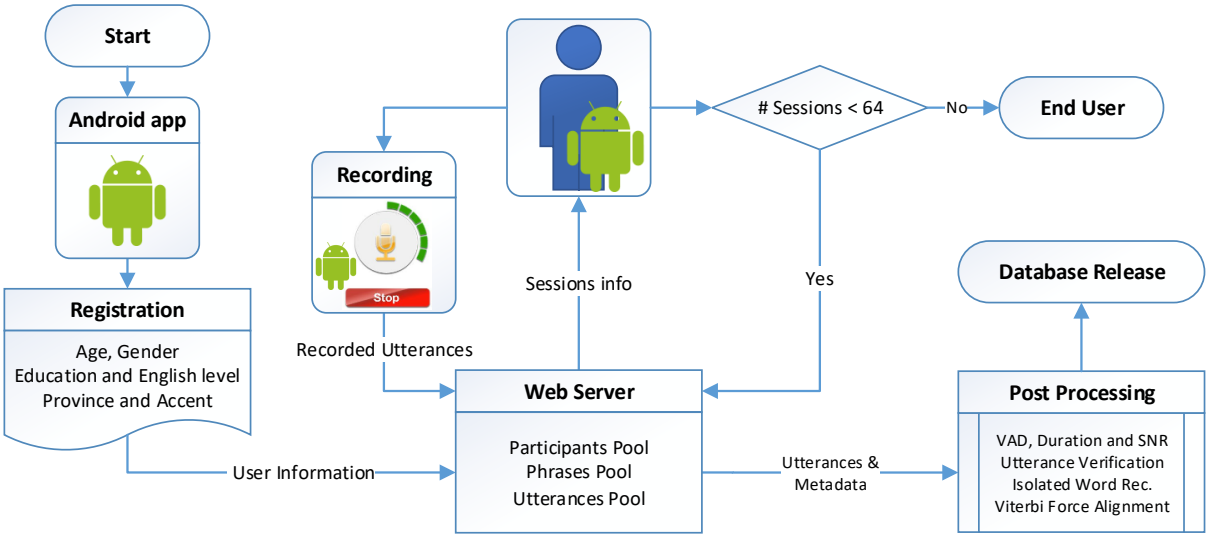

Figure 1: Flowchart of DeepMine data collection, from user registration to finishing 64 sessions.

4. Speech recognition and Viterbi force alignment will be performed on text-independent parts and the utterances will be grouped in several sets, based on their accuracy, similarly to LibriSpeech [13].

5. Several ASR methods will be performed to estimate the ground truth transcription. Furthermore, several of them will be checked by hand randomly to verify the estimated transcript. There is no plan to check the whole database manually because it is a very expensive task for such a large database.

3. The DeepMine Database Description

3.1. Data collection assumptions

As mentioned above, the goal of the DeepMine project is to create a public speech database with a large number of speakers which can be used in several speech processing tasks. Additionally, the following constrains are posed in the database design:

1. Each participant can contribute up to 64 sessions. Large number of sessions yields large number of target trials, leading to more valid evaluation results.

2. We avoid using too many phrases in each session, as this can discourage respondents from participating. For this purpose, similarly to the RedDots project, only 24 phrases are considered.

3. In order to reduce the overall contribution period for each participant, the minimum duration between two consecutive session recordings is set to 8 hours. This number is selected because we believe longer periods can make participants lose their interest in the project.

4. To further motivate participants, we award them with a number of prizes in lotteries. Lotteries are taking in a monthly basis, while the requirement to include a participant in the lottery is to complete at least 16 recording sessions.

Table 1: The number of different phrases in the current version of the DeepMine database. Note that these numbers may increase in the future.

	Count
Persian text-dependent	5
English text-dependent	5
Persian months-name, 12-months	10,000
Persian months-name, 3-months	1,320
English digits, 10-digits	5,000
English digits, 4-digits	5,040
Persian transcribed phrases	129,070

3.2. DeepMine Database Parts

The DeepMine database consists of three parts. The first part contains fixed common phrases to perform text-dependent speaker verification. The second part consists of random sequences of words useful for text-prompted speaker verification, and the last part includes phrases with phonetic level transcription, useful for text-independent speaker verification or speaker verification using a random phrase (similar to Part4 of RedDots). This part can also serve for Persian speech recognition. Each part is described in more details below. Table 1 shows the number of phrases in each part of the database.

1. **Part1 - Text-dependent:** This part contains a set of fixed phrases which are used to verify speakers in text-dependent mode. Each speaker utters 5 Persian phrases. In addition to these phrases, if the speaker can read English, five phrases selected from Part1 of the RedDots database are also recorded by the respondent. We choose the most simple phrases to help participants pronounce them correctly. The respondents can listen the correct pronunciation of the phrases, to help them pronounce them correctly.

2. **Part2 - Text-prompted:** In this part, 3 random sequences of

Persian month names are shown to the respondent. In [14], we showed that the performance on month names is better than the digits, and therefore they are also used here. These sequences are shown to the respondent in two modes. In the first mode, the sequence consists of whole 12 months which will be used for speaker enrollment. The second mode contains a sequence of 3 month names that will be used for evaluation. In each of the 8 sessions received from the server, there are 3 phrases of all 12 months (i.e. just in one session), and 7×3 other phrases for evaluation, containing fewer words. For a respondent who can read English, 3 random sequences of English digits are also recorded in each session. In one of the sessions, these sequences contain all digits and the remaining ones contain only 4 digits. Note that each word is unique in each random sequence. English digits are chosen because the correct pronunciation of English months can be difficult for many Iranian participants.

3. **Part3 - Text-independent:** In this part, 8 Persian phrases that have already been transcribed in phone level are displayed to the respondent. These phrases are chosen mostly from news. The minimum and maximum length for each one of them are also considered based on the number of phonemes they contain and their average duration, to prevent from having too short and too long recording. In this category, it is possible for the respondents to bypass a phrase without recording it, which may happen due to several reasons, such as mistakes in the phrase and inappropriate expressions. If the respondent is unable to read English, instead of 5 fix phrases and 3 random digit strings, 8 other Persian phrases are also prompted to the respondent, including 24 phrases in each recording session.

   This part has three usages. First, it can be used for text-independent speaker verification. This section contains a large number of speakers, phrases and utterances, making it appropriate for research on this task. Moreover, the phrases that are used in this part have different lengths, which can be useful for studying the effects of short duration on the performance of speaker verification systems. The second application of this part (similarly to Part4 of RedDots) is text-prompted speaker verification using random text (instead of a random sequence of words). Finally, the third application of this part is large vocabulary speech recognition in Persian.

   There is no proper database for speech recognition in Persian, apart from some limited cases which are collected in laboratory conditions [15]. Hence, this part can at least partly address this problem and enable robust speech recognition applications in Persian. Additionally, it can be used for speaker recognition applications, such as training deep neural networks (DNNs) for extracting bottleneck features [16, 17], as well as for extracting sufficient statistics using DNNs for i-vector training. The used phrases in this part are randomly selected from a collection of about 130,000 phrases.

3.3. Current State of the Database

After passing about two-thirds of the project's duration, 1355 respondents have been involved, with 775 of them being male and 580 female. 166 of them could not read English and are therefore participating only in Persian phrases. About 8900 sessions were recorded by females and similarly, about 6100 sessions by males, i.e. women are overrepresented in terms of sessions, even though their number is 14 percent lower than that of males. Other useful statistics related to the database are shown in Table 2.

Table 2: Different statistics of the DeepMine database in the current version.

	Count
Number of finished sessions	14954
Number of finished sessions contain English	13459
Number of recorded utterances	363300
Number of respondent with at least 5 utterances	1355
Number of respondent with at least 1 sessions	1236
Number of respondent with at least 4 sessions	740
Number of respondent with at least 8 sessions	538
Number of respondent with at least 16 sessions	393
Number of respondent with at least 60 sessions	54

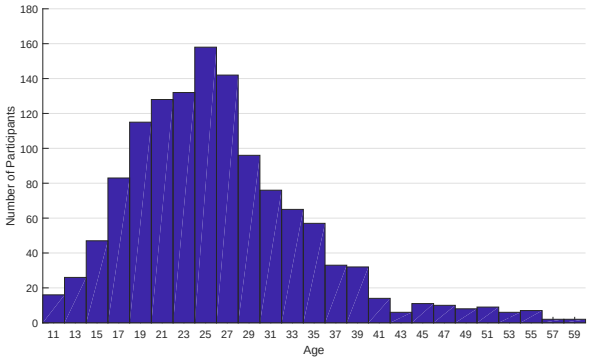

Figure 2: Age histogram of participants in DeepMine database.

One of the advantages of DeepMine database is the good coverage of different age groups. Figure 2 depicts the histogram of the age of the participants. Also, Figures 3 and 4 illustrate the educational and English level of participants.

The last status of the DeepMine database, as well as other related and useful information about its availability can be found on its website, together with a limited number of samples[1].

4. Preliminary Experiments and Results

We present here some preliminary results on speaker verification and speech recognition.

4.1. Speaker Verification Experiments

For evaluating the database on speaker verification, the i-vector based text-dependent method proposed in [18, 19] is applied on the male part of the database. In this experiment, 60-dimensional MFCC features are extracted from 16 kHz signal using HTK [20] with 25 ms Hamming windowed frames with 15 ms overlap.

The reported results are obtained with an i-vector based system. The 200-dimensional gender dependent i-vectors are length-normalized [21], and are further normalized using phrase dependent regularized LDA [19]. Cosine distance is used to obtain speaker verification scores. For aligning speech frames to Gaussian components, monophone HMMs with 3 states and 8 Gaussian components in each state are used [19].

[1] http://data.deepmine.ir/en/

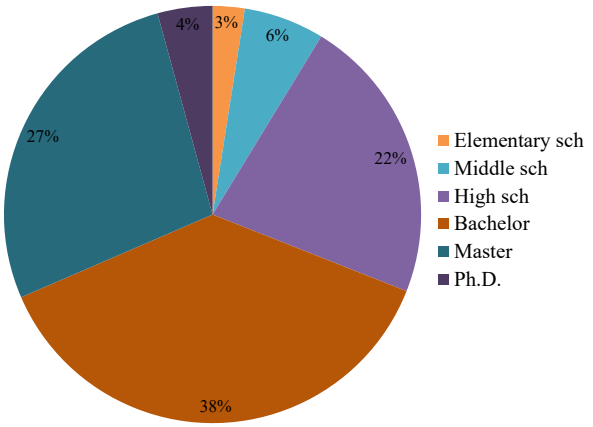

Figure 3: The chart of educational level of participant in the DeepMine database.

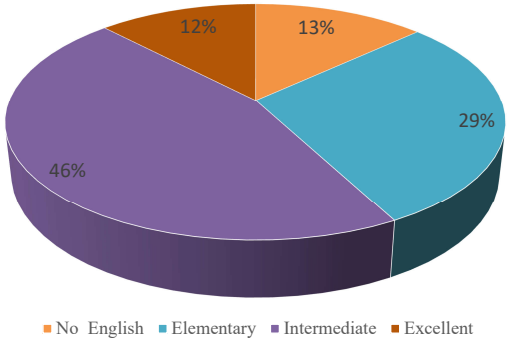

Figure 4: The chart of English level of participant in the DeepMine database.

Table 3 shows the results of text-dependent experiments and similarly Figure 5 shows the corresponding DET curves. For evaluating systems and plotting the DET curves, only the correct trials (i.e. Target-Correct vs Imposter-Correct) are considered as it has been proved that these are the most challenging trials in text-dependent speaker verification [17, 22].

We divided the database into three disjoint parts (training, development, and evaluation) as follows. Respondents with at least 21 recording sessions were selected as evaluation part. Similarly, respondents who had between 17 and 20 recording sessions were considered as development part and respondents with 2 to 16 recording sessions were included in the training set. Finally, respondents with one recording session or less were considered as unseen importers whose utterances should be evaluated against all speakers in the evaluation set.

Based on the above splits, there are 209 speakers in the evaluation set, 73 speakers in the development set and 660 speakers in the training set. In the experiments, the training set was used for HMM and i-vector training as well as for estimating the Regularized LDA (RLDA) transformation matrix, while the evaluation set was used to evaluate the method.

By comparing the results on Table 3 with our previous results on RSR2015 and RedDots databases, it is clear that the database is more challenging than RSR2015 and of similar difficulty with the RedDots database [19, 17, 22]. Same as RedDots, the female part of the database is more difficult than male part [22].

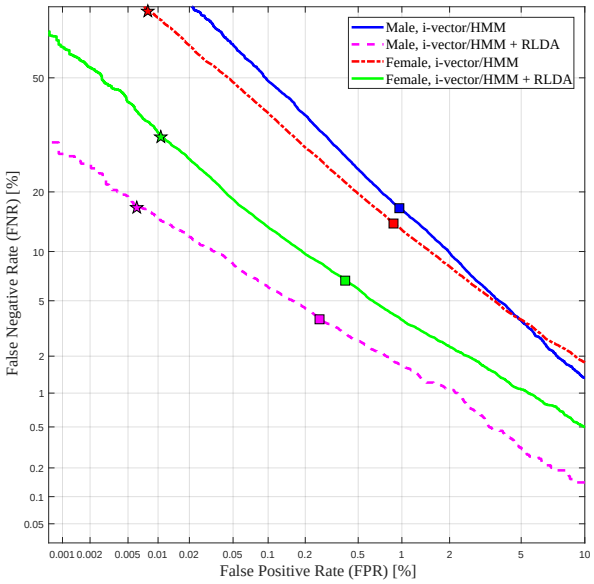

Figure 5: DET curves of two evaluated method on text-dependent part (i.e. Part1) of DeepMine database on Persian language. The square and star markers correspond to $\text{NDCF}^{\text{min}}_{\text{old}}$ and $\text{NDCF}^{\text{min}}_{\text{new}}$ operating points, respectively.

Table 3: i-vector/HMM method results on text-dependent part (i.e. Part1) of DeepMine database for males / females on Persian language.

Method	EER [%]	$\text{NDCF}^{\text{min}}_{\text{old}}$	$\text{NDCF}^{\text{min}}_{\text{new}}$
i-vector/HMM	4.35 / 4.24	0.263 / 0.228	0.865 / 0.767
i-vector/HMM + RLDA	1.33 / 2.20	0.063 / 0.107	0.230 / 0.437

4.2. Speech Recognition Experiments

In addition to speaker verification, we present several speech recognition experiments on Part3. The experiments were performed with the Kaldi toolkit [23]. For training HMM-based methods, about 80 hours of speech data were selected from 320 speakers and about 20 hours from 80 speakers were used for evaluation. A simple trigram language model is used for rescoring and the size of the dictionary is 90,000 words. Table 4 shows the results in terms of word error rate (WER) for different evaluated methods.

5. Conclusions

In this paper, we introduced a new speech database called DeepMine. We aim to collect a large-scale speech database through crowdsourcing which can be used for several speech processing tasks. Its main usage is the study of speaker recognition approaches in different short duration scenarios, such as text-dependent and text-prompted speaker verification. Additionally, it can be used for text-independent speaker verification with test segments of various durations. The third part of this database can be used for speech recognition, as it contains automatically annotated speech from random phrases with high

Table 4: WER of different methods on DeepMine database.

	WER [%]
MonoPhone	53.29
TriPhone + Deltas + Delta-Deltas	25.94
TriPhone + LDA + MLLT	24.07
TriPhone + LDA + MLLT + SAT	20.68
SGMM	17.52
MMI + SGMM	16.09

phonetic variability.

DeepMine database has several appealing features. First of all, it is a bilingual text-dependent speech database collected under non-laboratory environments. Therefore, the recordings contain considerable noise of various kinds. Thanks to the cooperation of the Iranian people to the database, it will be the largest publicly available database for text-dependent speaker recognition, which makes it suitable for research e.g. on deep learning models. The third part of the database consists of Persian automatically transcribed phrases, useful e.g. for training context dependent DNNs (e.g. bottleneck features) and evaluate them on other parts of the database. Moreover, it can be used for speech recognition, as our extensive experimentation demonstrates.

As of the time of this writing, DeepMine database contains more than 1300 speakers and about 360,000 utterances. About 400 of the participants have recorded at least 16 sessions and our goal is to reach at least 1000 speakers with minimum of 16 sessions. The first release of the dataset is planned for the fourth quarter of 2018.

6. Acknowledgments

The project was mainly supported by Sharif DeepMine company, a recently founded company in Iran. Also, the project was partially supported by ASR Gooyesh Pardaz (AGP) company, which develops speech processing systems in Persian.

7. References

[1] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms, Tech. Report CRIM-06/08-13," 2005.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] Y. Lei, N. Scheffer, L. Ferrar, and M. McLaren, "A novel scheme for speaker recognition using a phonetically aware deep neural network," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 1695–1699.

[4] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[5] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Joensuu, Finland, June 2014, pp. 1–8.

[6] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-End text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2016, pp. 5115–5119.

[7] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[8] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[9] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al., "The RedDots data collection for speaker recognition," in *InterSpeech*, 2015, pp. 2996–3000.

[10] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[11] Nic J De Vries, Jaco Badenhorst, Marelie H Davel, Etienne Barnard, and Alta De Waal, "Woefzela-an open-source platform for asr data collection in the developing world," 2011.

[12] Nic J De Vries, Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal, "A smartphone-based asr data collection tool for under-resourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2015, pp. 5206–5210.

[14] Hossein Zeinali, Elaheh Kalantari, Hossein Sameti, and Hossein Hadian, "Telephony text-prompted speaker verification using i-vector representation," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2015, pp. 4839–4843.

[15] M Bijankhan, J Sheikhzadegan, and MR Roohani, "Farsdat-the speech database of farsi spoken language," in *Australian Conference on Speech Science and Technology*, 1994.

[16] Fred Richardson, Douglas A. Reynolds, and Najim Dehak, "A unified deep neural network for speaker and language recognition," in *InterSpeech*, 2015, pp. 1146–1150.

[17] Hossein Zeinali, Hossein Sameti, Lukáš Burget, and Jan Černockỳ, "Text-dependent speaker verification based on i-vectors, deep neural networks and hidden Markov models," *Computer Speech & Language*, vol. 46, pp. 53–71, 2017.

[18] Hossein Zeinali, Lukas Burget, Hossein Sameti, Ondrej Glembek, and Oldrich Plchot, "Deep neural networks and hidden Markov models in i-vector-based text-dependent speaker verification," in *Odyssey-The Speaker and Language Recognition Workshop*, 2016, pp. 24–30.

[19] Hossein Zeinali, Hossein Sameti, and Lukas Burget, "HMM-based phrase-independent i-vector extractor for text-dependent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1435, 2017.

[20] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al., *The HTK book*, vol. 2, Entropic Cambridge Research Laboratory Cambridge, 1997.

[21] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *InterSpeech*, 2011, pp. 249–252.

[22] Hossein Zeinali, Hossein Sameti, Lukas Burget, Jan Cernocky, Nooshin Maghsoodi, and Pavel Matejka, "i-vector/HMM based text-dependent speaker verification system for RedDots challenge," in *InterSpeech*, 2016, pp. 440–444.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.