HUMAN AND MACHINE SPEAKER RECOGNITION BASED ON SHORT TRIVIAL EVENTS

*Miao Zhang, Xiaofei Kang, Yanqing Wang, Lantian Li, Zhiyuan Tang, Haisheng Dai, Dong Wang**

CSLT, Tsinghua University

ABSTRACT

Trivial events are ubiquitous in human to human conversations, e.g., cough, laugh and ahem. Compared to regular speech, these trivial events are usually short and unclear, thus generally regarded as not speaker discriminative and so are largely ignored by present speaker recognition research. However, these trivial events are highly valuable in some particular circumstances such as forensic examination, as they are less subjected to intentional change, so can be used to discover the genuine speaker from disguised speech.

In this paper, we collected a trivial event speech database that involves 75 speakers and 6 types of events, and report preliminary speaker recognition results on this database, by both human listeners and machines. Particularly, the deep feature learning technique recently proposed by our group is utilized to analyze and recognize the trivial events, which leads to acceptable EERs despite the extremely short durations (0.2-0.5 seconds) of these events. Comparing different types of events, 'Hmm' seems more speaker discriminative.

Index Terms— speaker recognition, speech perception, deep neural network, speaker feature learning

1. INTRODUCTION

Biometric authentication is highly important for the security of both reality and cyberspace. Among various biometrics, such as iris, palmprint, fingerprint and face, voiceprint has received much attention recently, partly due to its convenience and non-intrusiveness. After decades of research, speaker recognition (SRE) by voiceprint has achieved remarkable improvement [1, 2, 3, 4].

Most of the present SRE research works on 'regular speech', i.e., speech intentionally produced by people and involving clear linguistic content. For this type of speech, rich speaker information can be obtained from both vocal fold vibration and vocal tract modulation, so the speaker identifiablility is generally accepted. Many algorithms have been proposed to perform SRE with this kind of speech, including the statistical model approach that has gained the most popularity [5, 6, 7] and the neural model approach that emerged recently and has attracted much interest [8, 9, 10].

Despite the significant progress achieved on regular speech, research on non-linguistic part of speech signals is still very limited. For example, we may cough and laugh when talk to others, and may 'tsk-tsk' or 'ahem' when listening to others. These events are not intentionally produced by the speaker and contain little linguistic information. However, they do convey some information about the speaker. For example, we can recognize the person by even a laugh if we have been familiar with him/her. Because these non-linguistic, non-intentional events occur ubiquitously in our conversations, we call them 'trivial events'. Typical trivial events include cough, laugh, ahem, etc.

A key value of SRE on trivial events is that these events are resistant to potential disguise. In forensic examination, for example, the suspects may intentionally change their voices to counteract the voiceprint testing, which will largely fool the human listeners and fail the existing SRE system. However, trivial events are much harder to be counterfeited by the speaker, which makes it possible to use these events to discover the true speaker from disguised speech. We will show how disguised speech deceives humans and state-of-the-art SRE techniques in Section 5.

An interesting question is, which type of trivial events conveys more speaker information? Moreover, who are more apt to identifying speakers from these trivial events, humans or machines? In previous work, we have studied three trivial events: cough, laugh and 'wei' (Hello in Chinese), and found that with a convolution & time-delay deep neural network (CT-DNN), an unexpected high recognition accuracy can be obtained: the equal error rate (EER) reaches as low as 11% with a cough of 0.3 seconds [11]. This good performance is largely attributed to the deep speaker feature learning technique that we proposed recently [10].

In this paper, we extend the previous work [11] in several aspects: (1) we extend the study to 6 types of trivial events, i.e., cough, laugh, hmm, tsk-tsk, ahem and sniff; (2) we compare performance of human listeners and machines; (3) we collect a trivial event speech database and release it for public usage.

The organization of this paper is as follows: the deep feature learning approach is briefly described in Section 3, and then the trivial event speech database CSLT-TRIVIAL-I is presented in Section 4. The performance of human and ma-

This work was supported in part by the National Natural Science Foundation of China under Projects 61371136 and 61633013, and in part by the National Basic Research Program (973 Program) of China under Grant 2013CB329302. M. Zhang and X.F. Kang are joint first authors. Corresponding Author: Dong Wang.

chine tests is reported in Section 5, and some conclusions and discussions are presented in Section 6.

2. RELATED WORK

Speaker recognition on trivial events is still limited. The most relevant work we noticed is from Hansen et al. [12, 13]. They analyzed the acoustic properties of scream speech and studied the SRE performance on this type of speech using a recognition system based on the Gaussian mixture models-universal background model. Significant performance reduction was reported compared with the performance on regular speech.

Some studies are not for trivial speech events, but are still related to our work. For example, Fan et al. [14] investigated the impact of whispered speech on SRE, and Hanili [15] investigated the impact of loud speech.

3. DEEP FEATURE LEARNING

Most of existing speaker recognition techniques are based on statistical models, e.g., the Gaussian mixture model-universal background model (GMM-UBM) framework [5] and the subsequent subspace models, such as the joint factor analysis approach [6] and the i-vector model [7, 16]. Additional gains have been obtained by discriminative models and various normalization techniques (e.g., the SVM model [17] and PLDA [18]). A shared property of these statistical methods is that they use raw acoustic features, e.g., the popular Mel frequency cepstral coefficients (MFCC) feature, and rely on long speech segments to discover the distributional patterns of individual speakers. Since most of trivial events are short, these models are not very suitable to represent them.

The neural model approach gained much attention recently. Compared to the statistical model approach, the neural approach focuses on learning frame-level speaker features, hence more suitable for dealing with short speech segments, e.g., trivial events. This approach was first proposed by Ehsan et al. [8], where a vanilla deep neural network (DNN) was trained to discriminate the speakers in the training data, conditioned on the input speech frame. The frame-level features are then extracted from the last hidden layer, and utterance-based representations, called 'd-vectors', are derived by averaging the frame-level features. Recently, we proposed a new convolution & time-delay DNN (CT-DNN) structure, by which the quality of the learned speaker features is significantly improved [10]. Particularly, we found that the new features can achieve remarkable performance with short speech segments. This property has been employed to recognize two trivial events (cough and laugh) in our previous study, and good performance has been obtained [11]. More details about the CT-DNN model can be found in [10], including the architecture and the optimization method. The training recipe is also available online[1].

[1] http://project.cslt.org

In this paper, the deep feature learning approach will be heavily used to recognize and analyze more trivial events, and the performance will be compared with that obtained by human listeners.

4. DATABASE DESIGN

An appropriate speech corpus is the first concern before any analysis can be conducted on trivial speech events. Unfortunately, few trivial event databases are publicly available at present. The only exception is the UT-NonSpeech corpus that was collected for scream detection and recognition [13, 12], but this corpus contains only screams, coughs and whistles. As we are more interested in ubiquitous events that are not easy to be changed by speakers intentionally, a more complicated database is required. Therefore, we decided to construct our own database and release it for public usage. This database is denoted by CSLT-TRIVIAL-I.

To collect the data, we designed a mobile application and distributed it to people who agreed to participate. The application asked the participants to utter 6 types of trivial events in a random order, and each event occurred 10 times randomly. The random order ensures a reasonable variance of the recordings for each event. The sampling rate of the recordings was set to 16 kHz and the precision of the samples was 16 bits.

We received recordings from 300 participants. The age of the participants ranges from 20 to 60, and most of them are between 15 and 30. These recordings were manually checked, and those recordings with clear channel effect (noise, background babbling and echo) were deleted. Finally, the speech segments were purged and only a single event was retained (i.e., one cough or one laugh) in each segment. After this manual check, 75 persons were remained, with 5 to 10 segments for each event per person. Table 1 presents the data profile of the purged database.

Table 1. Data profile of CSLT-TRIVIAL-I

	Spks	Total Utts	Utts/Spk	Avg. dur
Cough	75	732	9.8	0.36
Laugh	75	709	9.5	0.39
Hmm	75	708	9.4	0.49
Tsk-tsk	75	1039	13.9	0.17
Ahem	75	691	9.2	0.45
Sniff	75	691	9.2	0.37

Besides the trivial event database, we also collected a disguise database. The goal of this data is to test how human listeners and the existing SRE techniques will be impacted by speakers' intentional disguise. This will provide a better understanding about the value of our study on trivial events.

The same application used for collecting CSLT-TRIVIAL-I was used to collect the recordings for the disguise database. Before the recording, the participants were instructed to try

their best to counterfeit their voices when recording the disguise speech. During the recording, the application asked the participants to pronounce 6 sentences. Each sentence was spoken twice, one time in the normal style and one time with intentional disguise. In manual check, segments with much channel effect were removed. After the manual check, recordings from 75 speakers were remained. This database is denoted by CSLT-DISGUISE-I. Table 2 presents the data profile in details.

CSLT-TRIVIAL-I and CSLT-DISGUISE-I have been released online[2]. Users can download them freely and use them under the Apache2.0 licence.

Table 2. Data profile of CSLT-DISGUISE-I

	Spks	Total Utts	Utts/Spk	Avg. dur
Normal	75	1202	5.3	2.28s
Disguised	75	1202	5.3	2.49s

5. EXPERIMENTS

This section reports our experiments. We first present some details of two SRE systems we built for the investigation, one based on the i-vector model and the other based on the deep speaker feature learning (denoted by d-vector system). Furthermore, performance with the two SRE systems on CSLT-TRIVIAL-I is reported and compared with the performance of human listeners. Finally, a disguise detection experiment conducted on CSLT-DISGUISE-I is reported, which demonstrates how speech disguise fools both humans and the existing SRE systems.

5.1. SRE systems

For the purpose of comparison, we build two SRE systems, an i-vector system and a d-vector system. For the i-vector system, the input feature involves 19-dimensional MFCCs plus the log energy, augmented by their and the first and second order derivatives. The UBM is composed of $2,048$ Gaussian components, and the dimensionality of the i-vector space is 400. Three scoring methods are used: cosine distance, cosine distance after LDA projection, and PLDA. The dimensionality of the LDA projection space is 150. When PLDA is used for scoring, the i-vectors are length normalized. The system is trained using the Kaldi SRE08 recipe [19].

For the d-vector system, the input feature involves 40-dimensional Fbanks. A symmetric 4-frame window is used to splice the neighboring frames, resulting in 9 frames in total. The number of output units is $5,000$, corresponding to the number of speakers in the training data. The frame-level speaker features are extracted from the last hidden layer, and the d-vector of each utterance is derived by averaging its frame-level speaker features. The scoring methods used for the i-vector system are also used for the d-vector system during the test, including cosine distance, LDA and PLDA.

The Fisher database is used as the training set, which was recorded by telephone and the sampling rate is 8 kHz. The training set consists of $2,500$ male and $2,500$ female speakers, with $95,167$ utterances randomly selected from the *Fisher* database, and each speaker has about 120 seconds of speech segments. This data set is used to train the UBM, the T matrix, and the LDA/PLDA models of the i-vector system, as well as the CT-DNN model of the d-vector system.

5.2. SRE on trivial events

In the first experiment, we evaluate the SRE performance on trivial events, by both human listeners and the two SRE systems. The CSLT-TRIVIAL-I database is used to conduct the test. It consists of 75 speakers and 6 types of trivial events, each involving 5-10 segments. As the original data of the recording is in 16 kHz, we down-sampled the signals to 8 kHz to match the Fisher database.

During the human test, the listener is presented 36 YES/NO questions, 6 questions per event type. For each question, the listener is asked to listen to two speech segments that are randomly sampled from the same event type, with a probability of 50% to be from the same speaker. Listeners are allowed to perform the test multiple times. We collected 33 test sessions, amounting to $1,188$ trials in total. The performance is evaluated in terms of detection error rate (DER), which is the proportion of the incorrect answers within the whole trials, thus including both false alarms and false rejections. The results are shown in Table 3. It can be seen that humans can tell the speaker from a very short trivial event, particularly with the nasal sound 'Hmm'. For Cough, Laugh, Ahem, humans can obtain some speaker information, but the performance is lower. For tsk-tsk and sniff, the performance is very bad, and the answers given by the listeners are almost random. This is expected to some extent, as these two types of events sound rather weak, and producing them does not use much of vocal fold and vocal tract.

Table 3. DER(%) of human test on trivial events.

DER%					
Cough	Laugh	Hmm	Tsk-tsk	Ahem	Sniff
20.20	20.71	19.7	42.42	26.26	35.86

For the machine test, there are about $260,000$ trials for each event type. The EER results with the i-vector system and the d-vector system are reported in Table 4. It can be observed that the d-vector system outperforms the i-vector system by a large margin, confirming that the deep speaker feature learning approach is more suitable than the statistical model approach when recognizing short speech segments. Comparing different events, it can be found that Hmm conveys the most speaker information, and Cough, Laugh, Ahem

[2]http://data.cslt.org

are less informative. Tsk-tsk is the least discriminative. All these observations are consistent with the results of the human test. The only exception is the performance on Sniff: with the i-vector system, it performs pretty bad, as in the human test; however with the d-vector system, the performance is significantly improved. The exact reason is still under investigation, but a hypothesis is that Sniff does involve speaker information, but as its signal is weak so the information is largely buried in channel noise, which causes difficulty for statistical models but impacts neural models less. This also suggests that human listeners may also rely on statistical clues to discriminate speakers, as the i-vector system does.

Comparing humans and machines, we can find that the best machine system, i.e., the d-vector system, is highly competitive. Although DER and EER values are not directly comparable, the results still show roughly that on almost all the types of trivial events, the d-vector system makes fewer mistakes than humans. Particularly on the events that humans perform the worst, i.e., Tsk-tsk and Sniff, machines work much better. Although the listeners we invited are not very professional, and the results may be impacted by the audio devices human listeners used, these results still provide a strong evidence that machines can potentially do better than human beings in listening trivial events.

Table 4. EER(%) results on CSLT-TRIVIAL-I with the i-vector and d-vector systems.

		EER%					
Systems	Metric	Cough	Laugh	Hmm	Tsk-tsk	Ahem	Sniff
i-vector	Cosine	23.42	27.69	15.71	29.70	18.12	37.78
	LDA	26.14	27.99	15.54	31.79	20.83	37.74
	PLDA	27.82	25.79	14.28	33.57	21.85	34.76
d-vector	Cosine	15.92	21.29	13.81	**27.30**	16.77	15.79
	LDA	18.69	21.28	13.69	28.94	17.08	17.49
	PLDA	**15.27**	**20.12**	**12.26**	27.77	**15.97**	**15.13**

5.3. Disguise detection

In the second experiment, we examine how humans and machines can discriminate disguised speech. For the human test, the listener is presented 6 trials, each containing two samples from the same speaker, but one of the sample can be a disguised version. The listener is asked to tell if the two samples are from the same speaker. To avoid any bias, the listeners are informed that some speech samples are disguised. Some trials may also involve imposter speech (not the same speaker), but these trial are only used to inject noise into the test, not counted in the final result. We collected 198 trails in total, and the DER result is 47.47%. This indicates that human listeners largely fail in discriminating disguised speech.

The EER results of the two SRE systems are reported in Table 5. It can be found that machines can do better than humans in discriminating disguised speech, but the error rates are still very high. Again, the d-vector system performs better than the i-vector system.

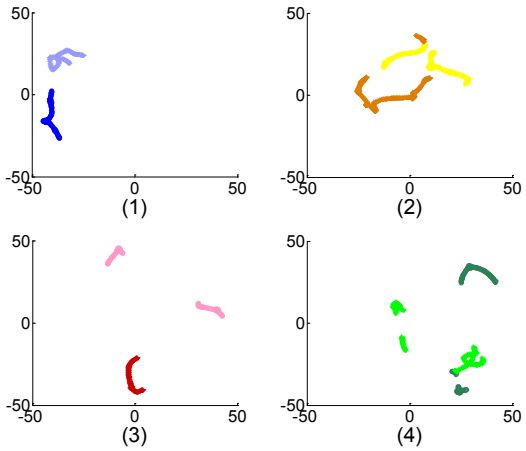

Fig. 1. The deep speaker features of the normal speech and disguised speech from the same speaker and the same sentence. Each picture represents a single person, and normal and disguised speech are represented by darker and lighter curves, respectively.

Table 5. EER(%) results on CSLT-DISGUISE-I with the i-vector and d-vector systems.

	EER%	
Metric	i-vector	d-vector
Cosine	28.70	25.74
LDA	34.57	**24.17**
PLDA	28.70	28.17

To observe the impact of speech disguise more intuitively, we plot the deep speaker features produced by the d-vector system in 2-dimensional space using t-SNE [20]. The results are shown in Fig. 1. We can see that the discrepancy between the normal and disguised speech is highly speaker-dependent: some speakers are not good voice counterfeiters, but some speakers can do it very well.

6. CONCLUSIONS

In this paper, we studied and compared the performance of human listeners and machines on the speaker recognition task with trivial speech events. Our experiments on 6 types of trivial events demonstrated that both humans and machines can discriminate speakers to some extent with trivial events, particularly those events involving clear vocal tract activities, e.g., 'Hmm'. Additionally, the deep speaker feature learning approach works much better than the conventional statistical model approach on this task, and in most cases outperforms human listeners. We also tested the performance of humans and machines on disguised speech, and found that speech disguise does place a serious challenge for both of them.

7. REFERENCES

[1] JA Unar, Woo Chaw Seng, and Almas Abbasi, "A review of biometric technology along with trends and prospects," *Pattern recognition*, vol. 47, no. 8, pp. 2673–2688, 2014.

[2] SR Kodituwakku, "Biometric authentication: A review," *International Journal of Trend in Research and Development*, vol. 2, no. 4, pp. 2394–9333, 2015.

[3] Homayoon Beigi, *Fundamentals of speaker recognition*, Springer Science & Business Media, 2011.

[4] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[5] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[7] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[8] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

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

[10] Lantian Li, Yixiang Chen, Ying Shi, Zhiyuan Tang, and Dong Wang, "Deep speaker feature learning for text-independent speaker verification," in *Proc. Interspeech 2017*, 2017, pp. 1542–1546.

[11] Miao Zhang, Yixiang Chen, Lantian Li, and Dong Wang, "Speaker recognition with cough, laugh and" wei"," *arXiv preprint arXiv:1706.07860*, 2017.

[12] John HL Hansen, Mahesh Kumar Nandwana, and Navid Shokouhi, "Analysis of human scream and its impact on text-independent speaker verification a," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. 2957–2967, 2017.

[13] Mahesh Kumar Nandwana and John HL Hansen, "Analysis and identification of human scream: Implications for speaker recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] Xing Fan and John HL Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[15] Cemal Hanilçi, Tomi Kinnunen, Rahim Saeidi, Jouni Pohjalainen, Paavo Alku, and Figen Ertas, "Speaker identification from shouted speech: Analysis and compensation," 2013.

[16] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[17] William M Campbell, Douglas E Sturim, and Douglas A Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.

[18] Sergey Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision–ECCV 2006*, pp. 531–542, 2006.

[19] 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, number EPFL-CONF-192584.

[20] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.