Speaker Recognition with Cough, Laugh and "Wei"

Miao Zhang*[†], Yixiang Chen*, Lantian Li* and Dong Wang*

* Center for Speech and Language Technologies (CSLT), RIIT, Tsinghua University
Tsinghua National Laboratory for Information Science and Technology
[†] Beijing University of Posts and Telecommunications

Abstract—This paper proposes a speaker recognition (SRE) task with trivial speech events, such as cough and laugh. These trivial events are ubiquitous in conversations and less subjected to intentional change, therefore offering valuable particularities to discover the genuine speaker from disguised speech. However, trivial events are often short and idiocratic in spectral patterns, making SRE extremely difficult. Fortunately, we found a very powerful deep feature learning structure that can extract highly speaker-sensitive features. By employing this tool, we studied the SRE performance on three types of trivial events: cough, laugh and "Wei" (a short Chinese "Hello"). The results show that there is rich speaker information within these trivial events, even for cough that is intuitively less speaker distinguishable. With the deep feature approach, the EER can reach 10%-14% with the three trivial events, despite their extremely short durations (0.2-1.0 seconds).

I. INTRODUCTION

Automatic speaker recognition (SRE) is an important biometric authentication technology. After several decades of development, SER has gained significant progress and the performance has been good enough for some constrained applications, e.g., with sufficient enrollment and test speech and the quality of the speech signals is reasonable [1], [2]. In spite of the prominent success, however, almost all the existing SRE approaches work on long-duration linguistic speech segments, e.g., segments with clear and long linguistic content such as "Hello, Google" [3]. In this paper, we are interested in some "trivial events" in speech signals, such as cough, laugh and "Wei" (a short "hello" in Chinese). These events are ubiquitous in conversations and often possess distinct properties, so it would be highly interesting to investigate how much speaker information is loaded in each of them. Moreover, SRE on trivial events may provide a powerful tool to attack speech disguising, as people who intend to disguise a personal identity are not easy to change her/his behaviors on these trivial events.

Very little has been done on these trivial events in SRE. This is not surprising: recognition on regular speech has been notoriously difficult, so it would be much harder to work with trivial events that are often short and may contain only non-linguistic content. Cough, for example, is often as short as 0.2 seconds, and the pronunciation is significantly different from regular speech: the airflow rushes out of the lung quickly and strongly vibrates the vocal cord. It is not easy to predict how much speaker information is involved in such a short and untypical signal, and it is even not easy to tell from our experience that if a speaker can be identified from a cough if we are not familiar with the person. Moreover, the training data for the trivial events are often very limited: although they are ubiquitous in conversations, the data volume takes only a very small proportion of the entire speech. We have not found any large-scale database that focuses on trivial events.

In this paper, we tackle the trivial event SRE problem by the deep speaker feature approach proposed recently by our group [4]. This approach designs a deep neural network (DNN) to learn frame-level speaker features from vast raw data, where the input of the model is a frame plus a short context, and the target involves all the speakers in the training database. The goal of the learning is to discover a speaker feature extractor (by the DNN hierarchy) that can extract speaker-sensitive features from a window of speech frames. Li et al. [4] reported very promising results and found that a good recognition accuracy can be obtained with even a very short speech segment. This feature learning approach therefore provides a powerful tool to discover the speaker information load of a small speech segment and retrieve it if it exists. We will use this tool to study the trivial events and try to answer the following three questions:

- Does a particular trivial event involve speaker information?
- Can the speaker information, if exists in a trivial event, be extracted from the event speech?
- Can the deep feature model trained with a regular speech database be migrated to recognize trivial event segments?

Our focus is put on three types of trivial events: cough, laugh and "Wei". We choose these types because they are among the most frequent ones in telephone conversations and are highly representative. First of all, cough is mostly related to vocal folder and involves little modulation from vocal track, so the spectrum contains no formants; laugh is also highly vocal-folder related, but the vocal tract may modulate the signal to some extent with the possible existed spectrum formant; "Wei" is mostly a regular speech segment, with a clear formant structure. Figure 1 displays the spectra of cough, laugh and "Wei" pronounced by the same speaker. It can be seen clearly that formant patterns exist in "Wei" but not clear in laugh, and totally absent in cough. From another perspective, cough is extremely short (less than 0.3 seconds), "Wei" is often very short (e.g., about 0.4 seconds), and laugh possesses a large variation in length, both intra-speaker and inter-speaker (e.g.,

from 0.3 to 1.0 seconds). Finally, the speaker information loads are intuitively different among these three types of events: it seems that "Wei" involves the most rich speaker information, as people can recognize who is speaking by a single "Wei" when picking up a phone. Laugh is the second and cough seems the mostly vague. In summary, investigation on the three types of events may provide a reasonable picture of the trivial event SRE.
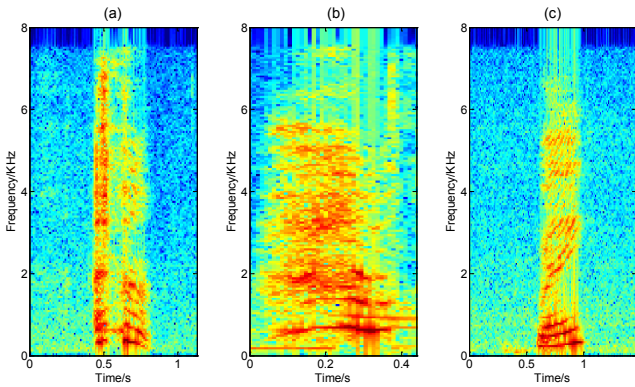

Fig. 1. Spectra of (a) Cough (b) Laugh (c) "Wei" spoken by the same person.

The organization of this paper is as follows: we firstly describe the deep feature learning approach in brief in Section II, and then present the trivial event speech database we collected for this study in Section III. The experiments will be presented in Section IV, followed by some conclusions and discussions in Section V.

II. DEEP FEATURE LEARNING

Most existing successful SRE approaches are model-based. For example, the famous Gaussian mixture model-universal background model (GMM-UBM) framework [5] and the subsequent subspace models, including the joint factor analysis approach [6] and the i-vector model [7]. They are generative models and heavily utilize unsupervised learning. Improvements have been achieved by either discriminative compensation (e.g., the SVM model [8] or PLDA [9]) or subspace modeling [10], [11]. Almost all these methods are based on raw acoustic features, e.g., the popular Mel frequency cepstral coefficients (MFCC) feature.

An advantage of model-based approaches is that they can discover the group-based behavior of speech signals of different speakers, and hence make appropriate decisions in the sense of maximum likelihood. However, the heavy reliance on statically models largely impedes the passion of researchers in discovering the inherent and essential mechanism that characterizes the traits of a speaker. Without knowing this essence, existing methods have to rely on a long speech segment to identify a speaker, by observing the 'distributional patterns' of the speech frames within the segment.

Using the model-based approach to discriminate trivial event speech segments would be difficult, due to the limited trivial event data for both training, enrollment and test. The only possible solution is to extract as much speaker information as possible from the short and untypical trivial event speech, and use a model as simple as possible to make the discrimination. This is the so called 'feature-based' approach. Unfortunately, traditional feature-based methods rely on human knowledge, which have been demonstrated to be ineffective even for regular speech, not to say the trivial event speech for which our knowledge is far from rich.

Fortunately, our recent research shows that it is possible to learn speaker sensitive features from raw speech signals by deep neural networks (DNN) [4], inspired by the seminar work of Ehsan [3]. We found a simple DNN model that can learn speaker features very well, and a good SRE performance can be obtained with a very small speech segment. This success in fact demonstrated that the speaker trait is largely a short-term spectral property, rather than a long-term distribution pattern. It also offers the possibility to discover the essential character of a speaker with very small trivial speech segments, for example, a cough or laugh.

The DNN structure we designed involves a few convolutional layers and several time-delayed layers: the former extracts local discriminative patterns, while the latter allows a long temporal context. This is referred as a CT-DNN model. Figure 2 illustrates the CT-DNN structure used in this work.

More specifically, the CT-DNN structure consists of a convolutional (CN) component and a time-delay (TD) component, connected by a bottleneck layer consisting of 512 hidden units. The convolutional component involves two CN layers, each followed by a max pooling. The TD component involves two TD layers, each followed by a P-norm layer. The settings for the two components, including the patch size, the number of feature maps, the time-delay window, the group size of the P-norm, have been shown in Figure 2. A simple calculation shows that with these settings, the size of the effective context window is 20 frames. The output of the P-norm layer is projected to a feature layer consisting of 400 units, which is connected to the output layer whose units correspond to the speakers in the training data.

This CT-DNN model can be trained easily by the natural stochastic gradient descent (NSGD) [12] algorithm. Once it has been trained, the speaker feature can be read from the feature layer, i.e., the last hidden layer of the model. As in [3], the utterance-level representation, called a 'd-vector', is derived by simply averaging the speaker features of all the frames within the utterance.

During test, the d-vectors of the enrollment and test utterances are produced respectively. The cosine distance between these two vectors is then used as the decision score for the SRE task. Similar to the i-vector system, some simple normalization methods can be employed to enhance the SRE performance, such as linear discriminant analysis (LDA) and probabilistic LDA (PLDA).

When applying the above deep feature learning approach to trivial event SRE, a particular problem is that the training data is highly sparse. It would be difficult to collect a large amount of trivial event speech segments, particularly cough and laugh. In this study, we simply employ a model that was trained
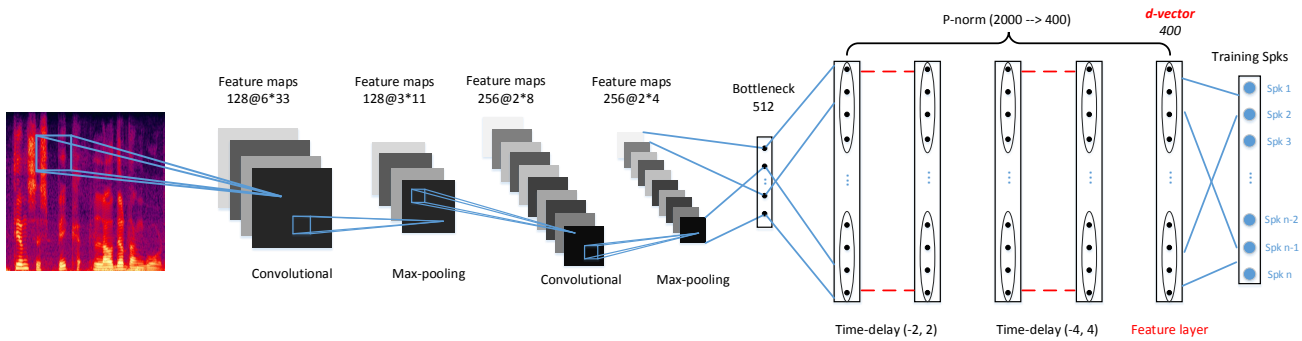
Fig. 2. The CT-DNN structure used for deep speaker feature learning.

with a large regular speech database. Our assumption is that the training data, although not aiming for trivial events, still involve some valuable information for them; particularly, some trivial event speech may be represented by regular phones, e.g., "Wei". This setting will also test the generalization capability of the deep feature learning model, and examine whether the speaker characterization produced from the trivial event speech is the same as that from the regular speech.

III. DATABASE CONSTRUCTION

Although the model can be trained with a regular speech database, the test data has to be specifically prepared. Since there is not any public database suitable for our study, we decided to construct a small trivial event speech database for the test, and release the data for public usage. This database is denoted by 'CSLT-COUGH100', and can be download online[1]. Table I presents the data profile in details.

TABLE I
DATA PROFILE OF CSLT-COUGH100

	Spks	Total Utts	Utts/Spk	Avg. dur
Cough	104	890	8.6	0.27s
Laugh	104	904	8.7	0.33s
"Wei"	104	848	8.2	0.37s

To collect the data, we designed a simple Android application, which instructs users to click a button to record cough, laugh and "Wei". The recording involved three sessions, each for one type of event. In each session, the speaker was instructed to pronounce the requested event (cough, laugh, "Wei") multiple times (not less than 8 times), with any variation they preferred. The recordings were then segmented into short segments by hand, each containing only a single event. The sampling rate of the recording is 16k Hz and the precision of the samples is 16 bits. The recording is mostly in the office environment, but some of them are collected on the street. The age of the participants ranges from 20 to 60, although most of them are between 20-30.

IV. EXPERIMENTS

This section reports our experimental results. We first describe the data and settings, and then report the SRE results

in terms of the equal error rate (EER). Some analysis for the discriminative power of the deep features are also presented.

A. Data

The Fisher database was used as the training set, which is recorded by telephone and the sampling rate is 8k Hz. The CSLT-COUGH100 was used as the test set. As the origin data of CSLT-COUGH100 is in 16k Hz, we down-sampled the signals to 8k Hz to match the Fisher database. More details of the two datasets are as follows.

- **Training set**: It 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 was 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.
- **Test set**: The CLST-COUGH100 database, consisting of 104 speakers. The database contains three types of trivial events (cough, laugh and "Wei"), and each type of event involves 5-10 segments. Details have been shown in Table I.

B. Model settings

For the purpose of comparison, an i-vector system was constructed as the baseline system. The raw features of this system involve 19-dimensional MFCCs plus the log energy. This raw features were augmented by the first and second order derivatives, resulting in feature vectors of 60 dimensions. The UBM was composed of $2,048$ Gaussian components, and the dimensionality of the i-vector space was 400. The dimensionality of the LDA projection space was set to 150. When using PLDA as the scoring metric, the i-vectors were length normalized. The system was trained using the Kaldi SRE08 recipe [13].

The d-vector system uses the CT-DNN architecture shown in Figure 2. The input features were 40-dimensional Fbanks. A symmetric 4-frame window was used to splice the neighboring frames, resulting in 9 frames in total. The number of output units was $5,000$, corresponding to the number of speakers in the training data. The frame-level speaker features were extracted from the last hidden layer (the feature layer in

---

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

Figure 2), and the utterance-level d-vectors were derived by averaging the frame-level speaker features. The scoring methods used for the i-vector system were also used for the d-vector system during the test, including cosine distance, LDA and PLDA.

C. Main results

The EER results of the i-vector system and the d-vector system are reported in Table II. It can be observed that with the best i-vector baseline (cosine scoring), the performance on "Wei" (12.72%) is reasonably good considering the short duration of the test utterance. On cough and laugh, the performance is significantly reduced (19.96% and 23.03% respectively). These results are expected, as the model are not intentionally trained to cover these two kinds of trivial events, and the content of these two events are largely non-linguistic, so likely involve less speaker information.

TABLE II
EER(%) RESULTS WITH THE I-VECTOR AND D-VECTOR SYSTEMS.

| Systems | Metric | EER% | | |
		Cough	Laugh	"Wei"
i-vector	Cosine	19.96	23.03	12.72
	LDA	23.55	24.24	12.90
	PLDA	23.33	24.30	13.77
d-vector	Cosine	11.19	13.62	10.66
	LDA	12.37	13.41	10.75
	PLDA	10.99	13.76	10.06

The d-vector system is significantly better than the i-vector system, demonstrating that the feature-based approach is more powerful. The best d-vector system is the one with the PLDA scoring, and the EERs with this system can reach 10.99% ,13.76% and 10.06% for cough, laugh and "Wei", respectively. The lower EER for "Wei" compared to the i-vector system is something expected, as we have demonstrated in [4] that the d-vector system is stronger than the i-vector system on small speech segments such as "Wei". The good performance with cough and laugh, however, is a bit surprising: both the two events do not contain linguistic content, but the performance is not significantly worse than the linguistic event "Wei". This seems that non-linguistic events still involve rich speaker information, and implies that our vocal cords are highly complex and speaker specific. The low performance with the i-vector model on the trivial events should not be explained as little speaker information embedded within these events; instead, it is just because the i-vector model cannot extract the information and utilize it well.

Comparing the results on cough and laugh, it can be seen that the performance on laugh is slightly worse than on cough. This is again a little unexpected. From our experience, it seems that we can recognize a person easier by her/his laugh than cough. A possible explanation is that the laugh speech may involve significant within-speaker variations, due to the freedom within the vocal tract modulation. The cough, on the other hand, is less modulated by the vocal tract and thus more stable. To have an intuitive comparison, Figure 3 shows the spectra of three cough segments and three laugh segments

from the same speaker. It can be seen that the three laugh segments are clearly different, while the cough segments are pretty much the same.

Finally, we found that for both two systems, the discriminative normalization approaches, LDA and PLDA, did not provide clear advantage. This is particularly the case for the i-vector system for which all the normalization methods reduce the performance on all the three types of trivial events. A possible reason is that trivial events are different from regular speech, so the LDA and PLDA models trained with the regular speech database are not very suitable. This argument explains why the performance reduction with the normalization methods is most significant on cough, if we admit that cough is the most distinct from regular speech among the three types of events.
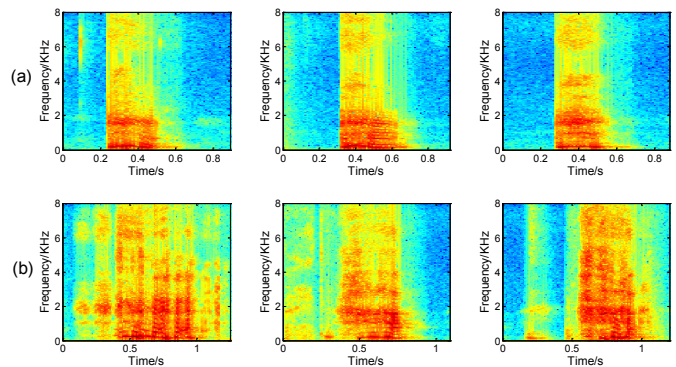

Fig. 3. The spectra of (a) Cough (b) Laugh segments of the same speaker.

D. Feature discrimination

In the last experiment, we investigate the distribution of the deep speaker features, particularly the speaker variations with these three types of trivial events. For this purpose, we randomly selected 10 speakers, and drew the speaker features belonging to these three types of events respectively using t-SNE [14]. The results are presented in Figure 4. It can be seen that the learned features with "Wei" are reasonably discriminative for speakers. But there are still variations appeared in cough and laugh figures as seen from the plot (a) and (b).

V. CONCLUSIONS

This paper employed the deep speaker feature learning approach to perform SRE on trivial events, and the focus was put on cough, laugh and "Wei". Our experiments showed that the deep feature model trained on a regular speech database (Fisher) can be used to perform trivial event SRE with an unexpected success. In spite of the extremely short duration, the EER can be as low as 10%-14%, depending on the type of events. These results can answer the questions raised in the introduction session:

- Does a particular trivial event involve speaker information? Yes. At least for the three trivial events studied in this paper, rich speaker information is involved.
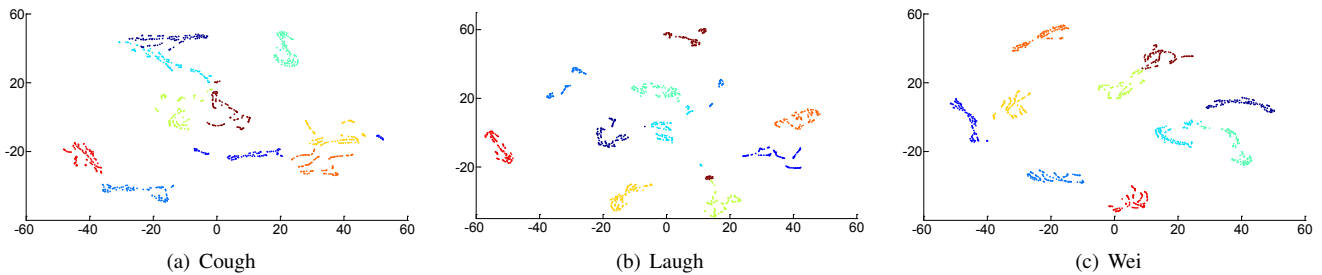
Fig. 4. Deep speaker features on events (a) Cough (b) Laugh (c) "Wei" randomly sampled from 10 speakers. The pictures are plotted by t-SNE, with each color representing a speaker.

- Can the speaker information, if exists in a trivial event, be extracted from the short segment? Yes. The deep feature approach was capable of extracting the speaker information from the short and idiocratic trivial events.
- Can the deep feature model trained with a regular speech database be migrated to recognize trivial event segments? Yes. A DNN model trained with the Fisher database worked well on trivial event SRE.

There is much work remaining: how about the performance on other trivial events, e.g., En, Ah, tapping and flapping speech? What is the implication of the experimental results for the acoustic and linguistic research? How the performance will be in a true scenario of speech disguise? All are under investigation.

REFERENCES

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

[2] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[3] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. 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.

[4] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," in *Interspeech 2017*, 2017.

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

[6] P. Kenny, G. Boulianne, P. Ouellet, and P. 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] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. 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] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

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

[10] 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*, 2014, pp. 293–298.

[11] Y. Lei, N. Scheffer, L. Ferrer, and M. 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.

[12] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of dnns with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.