HUMAN AND MACHINE SPEAKER RECOGNITION BASED ON SHORT TRIVIAL EVENTS

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

Center for Speech and Language Technologies, Tsinghua University

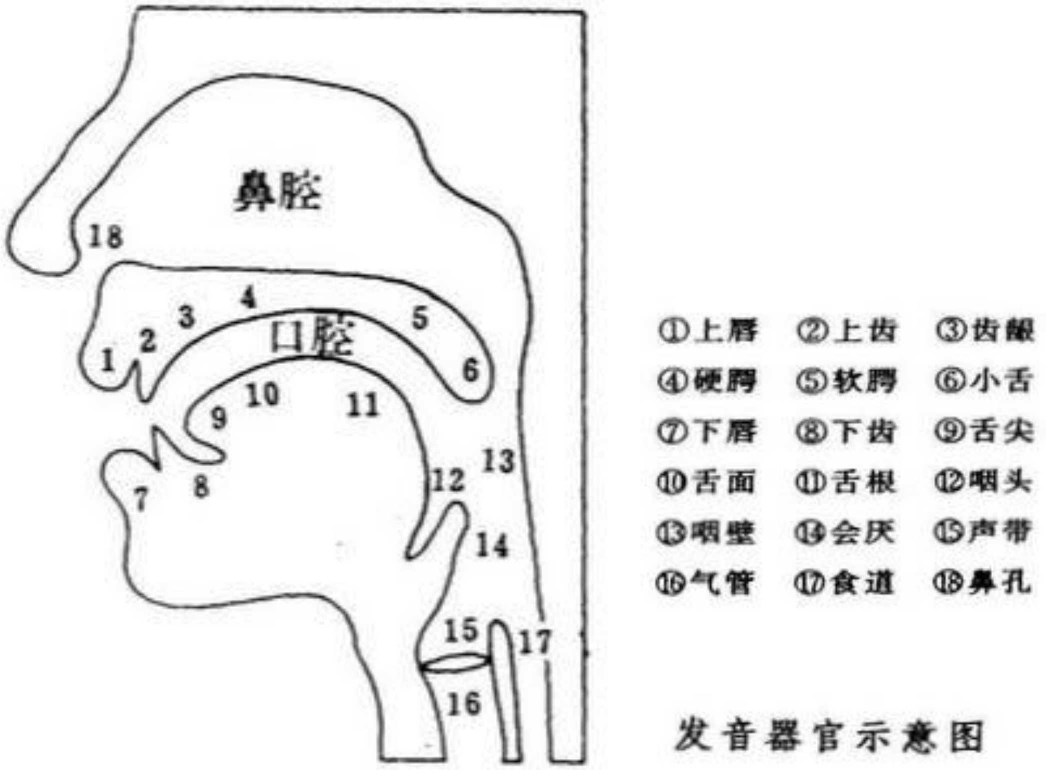
ICASSP, April 15-20, 2018, Calgary, Canada

Outline

- Background and significance
- Deep feature learning
- Databases and Experiments
- Results and discussions
- Further work

Background

- Normal sound
- Trivial sound
  - inspiratory sound
  - expiratory sound
    - vocal cord non-related
    - vocal cord related
      - Oral cavity resonance
      - nasal cavity resonance



发音器官示意图

Background

- Normal sound
- Trivial sound
  - inspiratory sound
  - expiratory sound
    - vocal cord non-related
    - vocal cord related
      - Oral cavity resonance
      - nasal cavity resonance

'Tsk-tsk' *Sniff*

Cough *'Ahem'*

Laugh

'Hmm'

Background

Most of the present SRE research works on ‘regular speech’, intentionally produced by people and involving clear linguistic content

+

Very little has been done on these trivial events in SRE (short duration & significantly different pronunciation & no large-scale specific database)

solution

1. A Trivial Events Database
2. A powerful tool to learn speaker feature

Significance

We have answered:

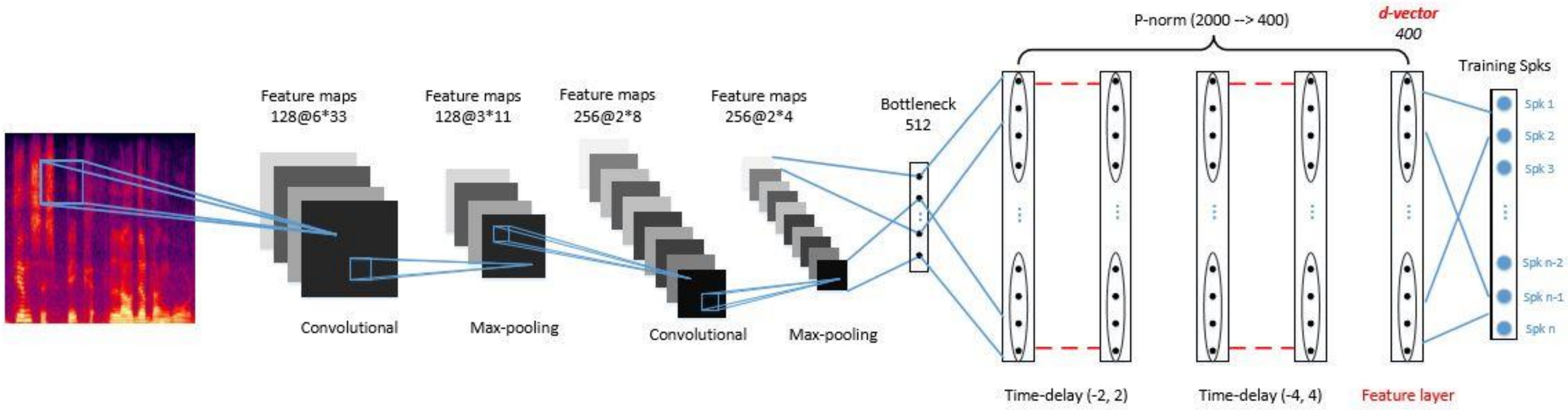
- Some particular trivial events do involve speaker information.
- The speaker information can be extracted from the trivial event speech.
- Deep feature model trained with a regular speech database can be migrated to recognize trivial event segments.

We want to explore:

- Which type of trivial event(phonation mechanism) conveys more speaker information?
- Who is more apt to identify speakers from these events, human or machine?
- Speaker recognition tasks on difficult situations, such as disguised speech.

Deep feature learning

CT-DNN model can learn speaker sensitive features, which is highly discriminative and can be used to achieve impressive performance when the test utterances are extremely short (0.2-0.5 seconds).



Databases

Participants utter 6 types of trivial Events in a random order, and each event occurred 10 times randomly.

Recordings from 75 persons were remained, with 5 to 10 segments for each event per person.

	Spks	Total Utts	Utts/Spk	Avg. duration (s)
Cough	75	732	9.76	0.36
Laugh	75	709	9.45	0.39
'Hmm'	75	708	9.44	0.49
'Tsk-tsk'	75	1039	13.85	0.17
'Ahem'	75	691	9.21	0.45
Sniff	75	691	9.21	0.37

Experiments

- Machine tests:

An i-vector system was constructed as the baseline system; A d-vector system uses the CT-DNN architecture.

- Human tests:

Listeners are 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, and they tell if the two samples are from the same speaker.

Result & Discussion

		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**

? D-vector system has general better performance than i-vector system.

! Deep speaker feature learning approach is more suitable than the statistical model approach on short speech segments.

Result & Discussion

		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	8.89	12.43	**5.88**	16.75	10.44	**11.91**
	LDA	**8.33**	**11.20**	6.76	**15.95**	**9.71**	12.44
	PLDA	10.26	15.48	7.28	17.85	13.16	12.93

? Machine performs best on 'hmm'.

! 'hmm' conveys the most speaker information.

- **Vocal cord & vocal track**
- **Resonation**

Result & Discussion

		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	8.89	12.43	**5.88**	16.75	10.44	**11.91**
	LDA	**8.33**	**11.20**	6.76	**15.95**	**9.71**	12.44
	PLDA	10.26	15.48	7.28	17.85	13.16	12.93

? Machine performs well on cough, 'ahem' and laugh.

- **Vocal cord & vocal track**

! Cough, 'ahem', laugh are less informative than 'hmm'.

Result & Discussion

		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	8.89	12.43	**5.88**	16.75	10.44	**11.91**
	LDA	**8.33**	**11.20**	6.76	**15.95**	**9.71**	12.44
	PLDA	10.26	15.48	7.28	17.85	13.16	12.93

? Machine performs worst on 'sniff' and 'tsk-tsk'.

! 'Tsk-tsk' and Sniff are the least discriminative.

Result & Discussion

		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	8.89	12.43	**5.88**	16.75	10.44	**11.91**
	LDA	**8.33**	**11.20**	6.76	**15.95**	**9.71**	12.44
	PLDA	10.26	15.48	7.28	17.85	13.16	12.93

? LDA and PLDA did not provide clear advantage on 'hmm' and sniff.

! Little intra-speaker variances.

Human Test

DER%					
Cough	Laugh	‘Hmm’	‘Tsk-tsk’	‘Ahem’	Sniff
20.20	20.71	19.70	42.42	26.26	35.86

- Human test results are consistent with the machine test.

- On almost all the types of trivial events, the d-vector system makes fewer mistakes than humans.

Databases

Participants pronounce 6 sentences, each involving 5 to 10 words. Each sentence was spoken twice, one time in the normal style and one time with intentional disguise.

Recordings from 75 speakers were remained.

	Spks	Total Utts	Utts/Spk	Avg. duration (s)
Normal	75	410	5.47	2.28
Disguised	75	410	5.47	2.49

Disguise detection

Machine test

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

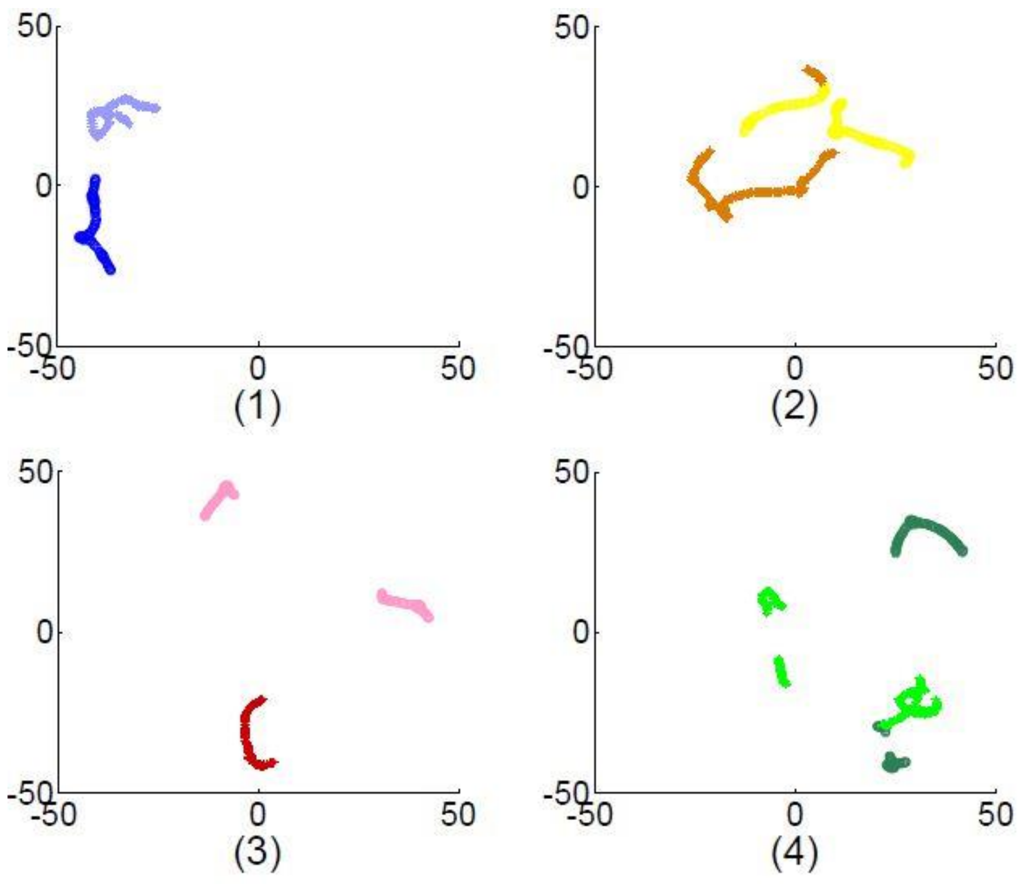
Human test

DER%: 47.47

- Machines can discriminate disguised speech to some extent, but the error rate Is much higher than that on normal speech.

- Again, the d-vector system performs better than the i-vector system.

Disguise detection



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.

Disguise speech impact

- Which type of trivial event(phonation mechanism) conveys more speaker information?

According to the six types of trivial events studied in this work, vocal cord and resonation related events convey more speaker information.

- Who is more apt to identify speakers from these events, human or machine?

Machine. And the deep feature learning system far outperforms the traditional i-vector system on short speech.

- Speaker recognition tasks on difficult situations, such as disguised speech?

Neither machine or human did well in discriminating disguised speech.

Summary

Text-independent	Habit-related	Subconscious
√	√	√

Further works

- Find out the way that the speaker information embed into phonation in order to improve the precision of trivial event recognition.

- Build up the SRE method based on trivial event and apply it to difficult scenarios, e.g., disguise scenario, time-varying scenario, emotion scenario.

- Explore how to combine the recognition results on normal speech and trivial speech, then improve the overall performance of speaker recognition system.

Thanks a lot.