Speaker Recognition with Cough, Laugh and “Wei”

Miao Zhang & Yixiang Chen

Center for Speech and Language Technologies
GROUPING

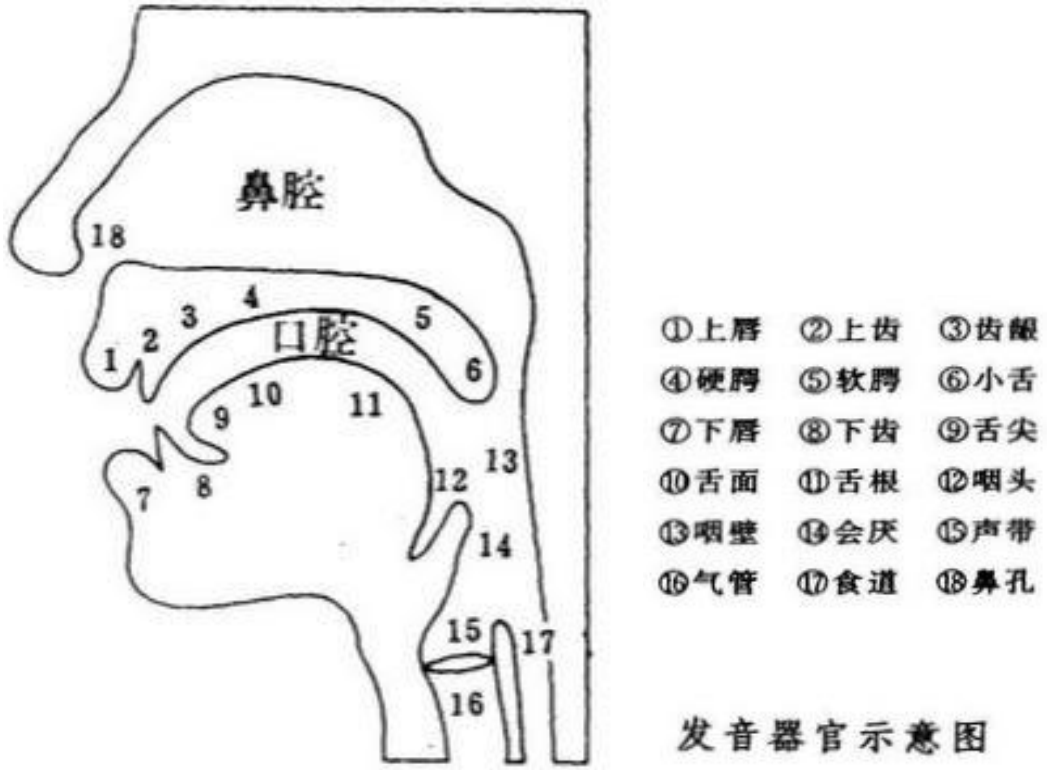
July 3, 2017

Outline

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

Background

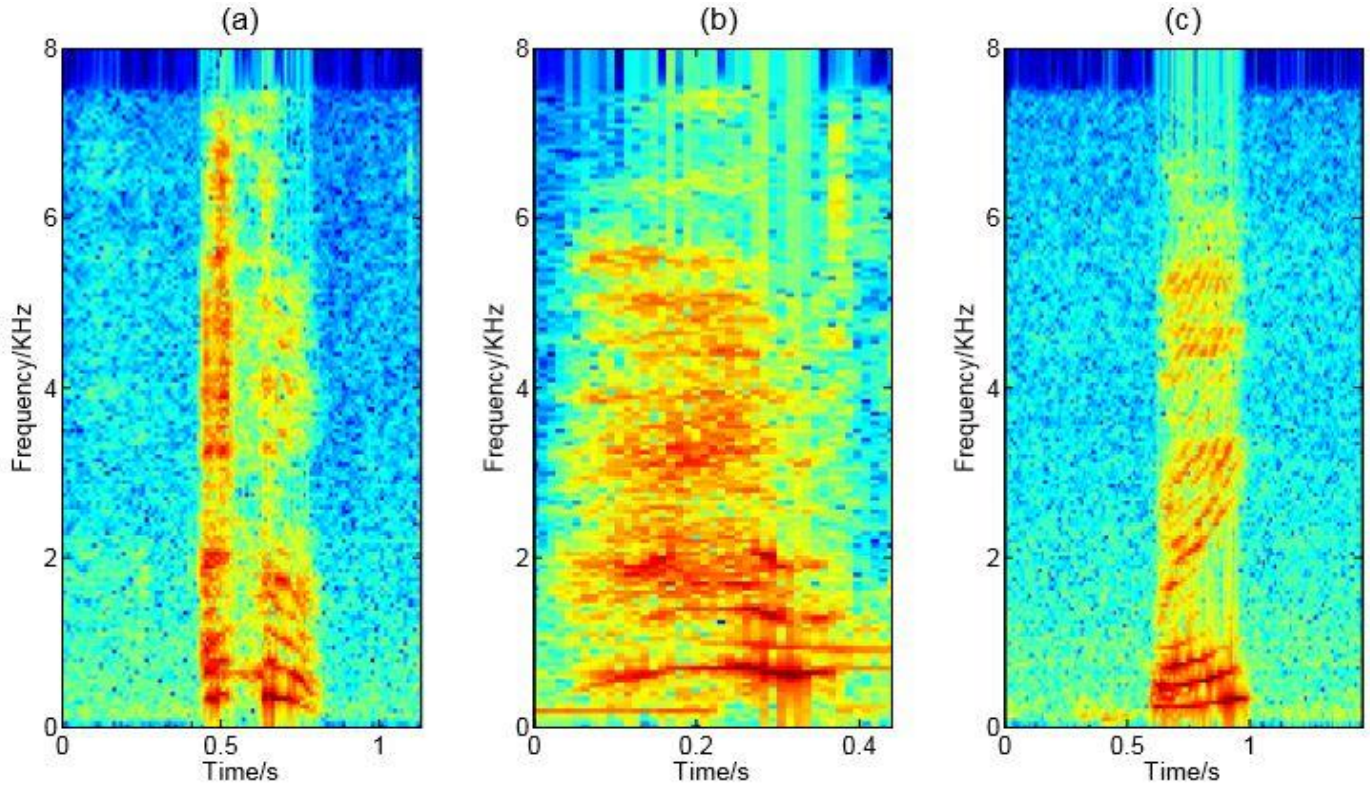
- Linguistic sound
- Natural sound
  - inspiratory sound
  - expiratory sound
    - vocal cord non-related
    - vocal cord related
      - Oral cavity resonance
      - nasal cavity resonance



发音器官示意图

Background

- Different spectra of trivial events cough, laugh and Wei



cough laugh Wei

Background

Almost all existing automatic speaker recognition(SRE) approaches work on long and clear linguistic content such as “Hello, Google”.

+

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

With the deep feature learning structure, strong speaker features can be learnt from a very short segment (about 0.3second)

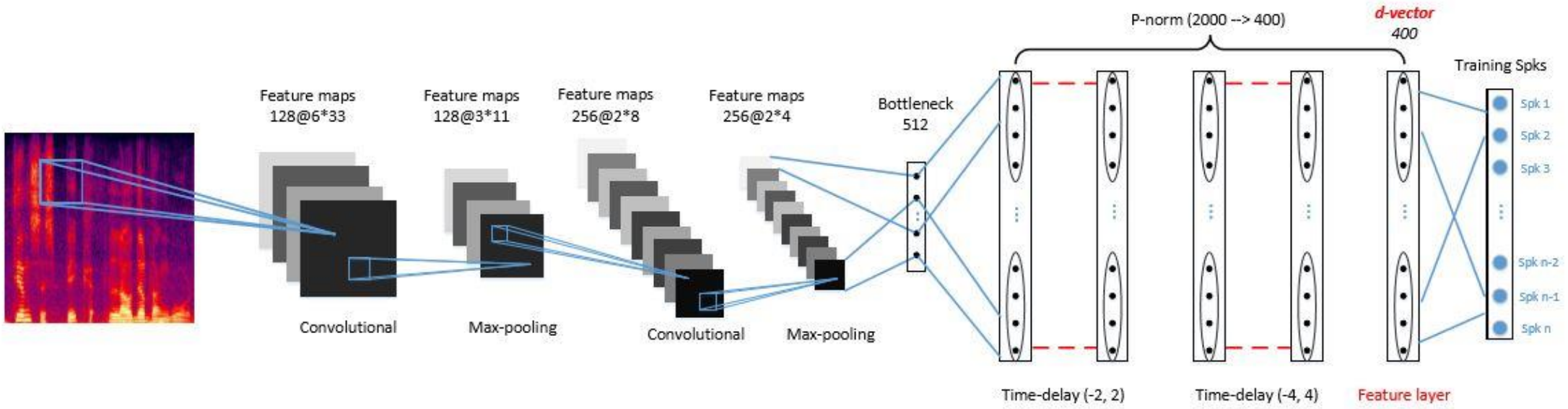
Significance

Answer 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?

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.



Experiment

- About database

Train with the *Fisher* speech database and construct the test database with recording.

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

- About data

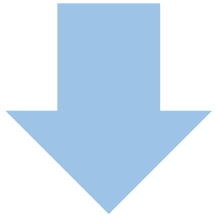
Every speaker has three types of events and every event contains 5—10 segments, with the most common 7—10.

Experiment

- An i-vector system was constructed as the baseline system
- The d-vector system uses the CT-DNN architecture
- Use t-SNE to investigate feature discrimination with the three types of events

Result & Discussion

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



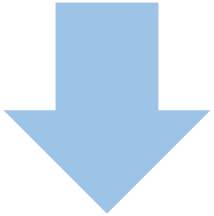
The feature-based approach is more powerful.

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

		EER%		
Systems	Metric	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

Result & Discussion

- For i-vector system, the performance on “Wei” is good, and the performance on cough and laugh is significantly reduced.



The i-vector model cannot extract the information in trivial events (cough and laugh) well

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

		EER%		
Systems	Metric	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

Result & Discussion

- For d-vector system, the best performance is on Wei, but the performance on cough and laugh is not significantly reduced.

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

		EER%		
Systems	Metric	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

1. Non-linguistic events still involve rich speaker information
2. The d-vector system could deal with them better.

Result & Discussion

- For d-vector system, the performance on laugh is slightly worse than on cough.

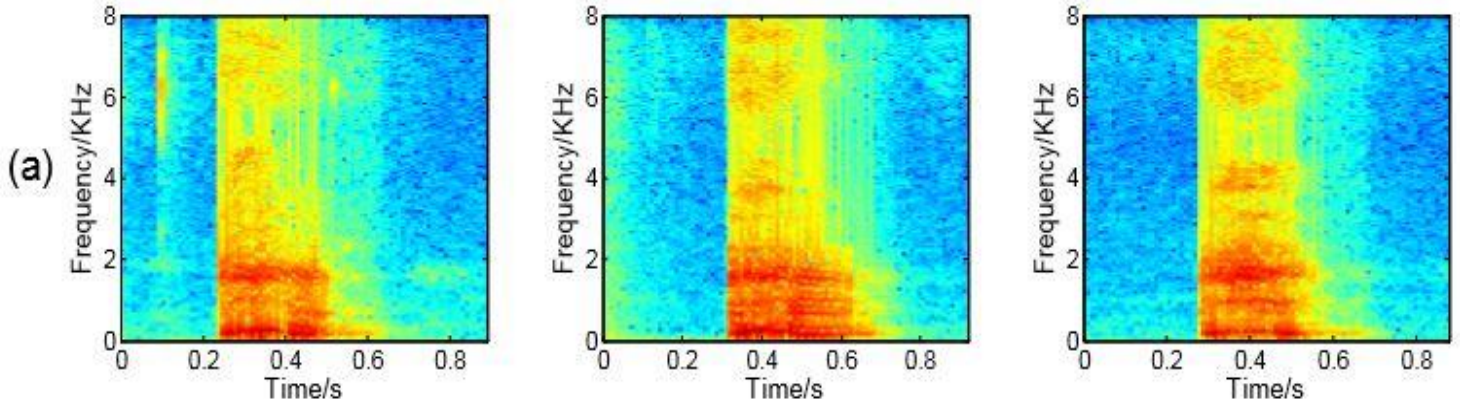
EER(%) RESULTS WITH THE I-VECTOR AND D-VECTOR SYSTEMS.

		EER%		
Systems	Metric	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

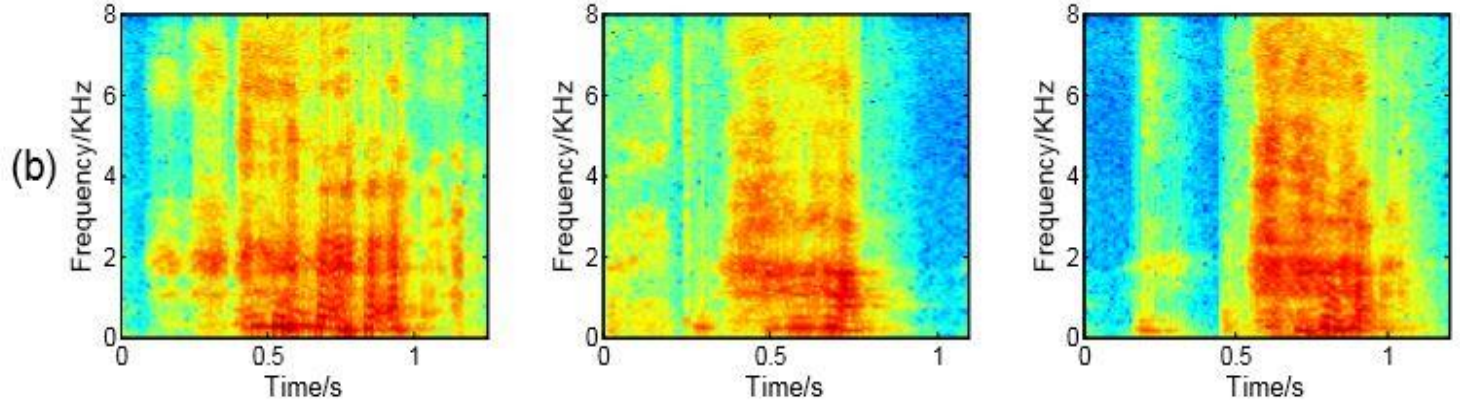
Laugh speech involve significant within-speaker variations related to vocal tract modulation.

Discussion

cough



laugh



Result

- For both two systems, the discriminative normalization approaches, LDA and PLDA, did not provide clear advantage.

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

		EER%		
Systems	Metric	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

LDA and PLDA models trained with the regular speech database are not very suitable.

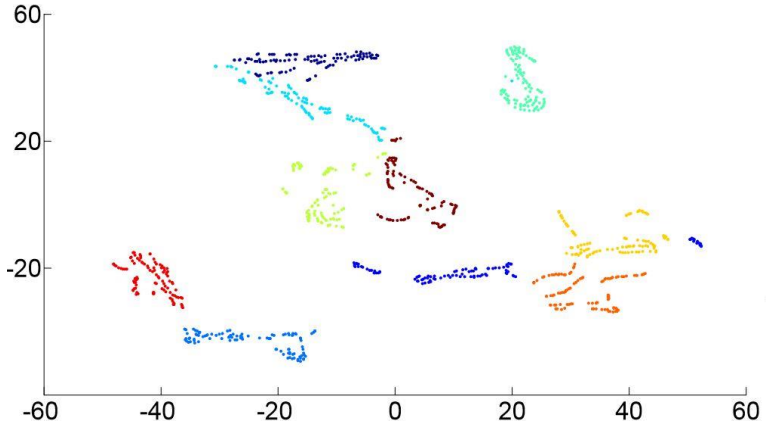
Result

- *Wei*

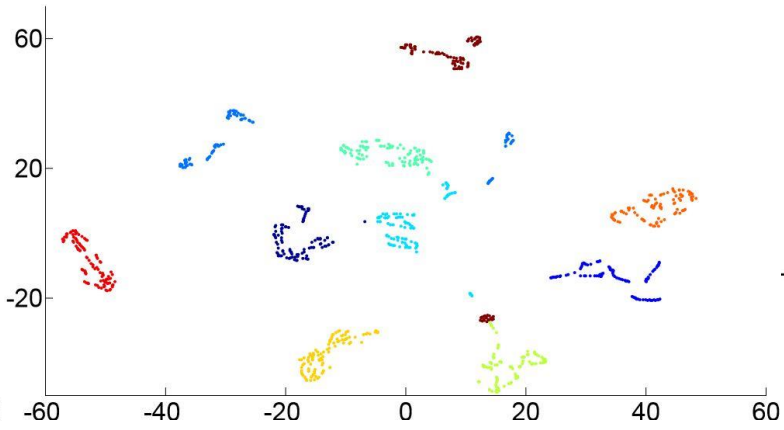
The learned features are reasonably discriminative for speakers.

- *cough and laugh*

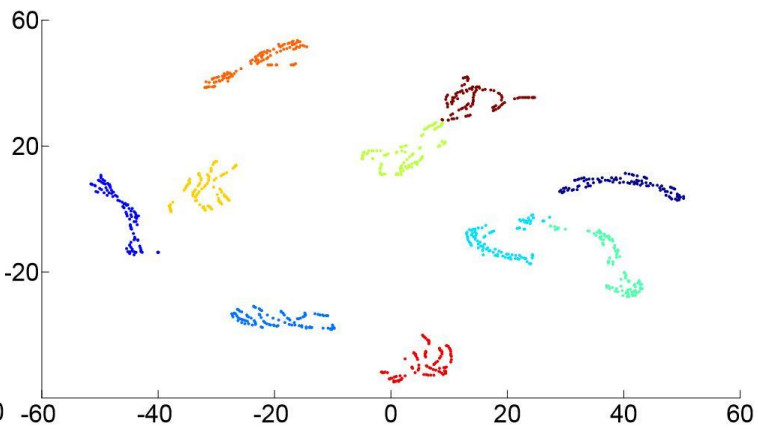
There are still variations.



cough



laugh



Wei

Conclusion

A satisfying performance!!

In spite of the extremely short duration and lack of linguistic information, the EER can be as low as 10%-14%, depending on the type of events.

Conclusion

- 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.

- 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.

Further work

- How about the performance on other trivial events (every type referred in background), and which one is most discriminative?
- 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?

Thank you