Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios

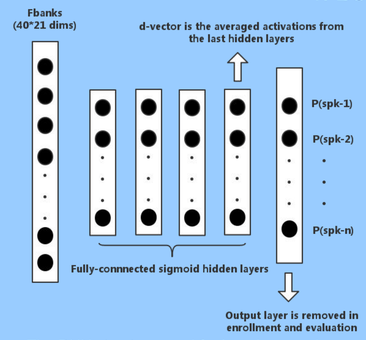
Renyu Wang, Mingliang Gu
School of Linguistic Science, Jiangsu Normal University, China
Lantian Li, Mingxing Xu, **Thoms Fang Zheng**
Center for Speech and Language Technologie, Tsinghua University, China

1 Motivation

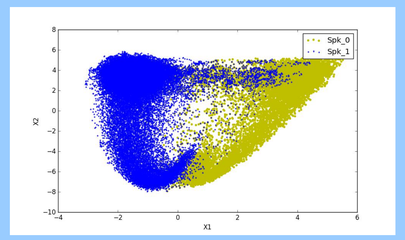
- Speaker segmentation is diffcult and many exsiting methods cannot work in fast speaker change scenarios.
- Common methods in metric-based segmentation to discriminate different speakers are based on some distance measure assumptions based on probabilistic models that require a certain length of voice to make the statistical result stable.
- If the analysis window size is too long, there might be more than one speaker change points in the two adjacent windows, if it is too short, speaker characteristics cannot be extracted accurately.
- Challenges:
  - Very diffcult to extract speaker discriminative feature in a short time window.
  - In short time, speaker characteristics are more sensitive to nuisance variation such as speech content and channel noise.

2 Methods

1. Deep speaker vectors



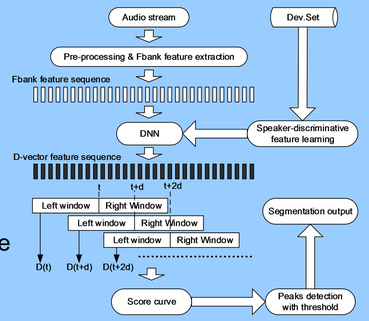
The DNN structure used for learning speaker-discriminative features



Plot of the d-vectors of two speakers on a 10-minutes conversation (with PCA dimensionality reduction to 2 dimensions)

2. Framework of segmentation

- Pre-processing and filter-bank feature extraction.
- Feed feature sequence to DNN to generate d-vector sequence.
- Calculate the distance between two adjacent windows with cosine distance of d-vectors and generate score curves.
- Change points often detected around the local minimum values.



The d-vector segmentation architecture

3 Results

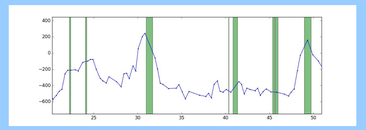
1. Speaker discriminative ability in short time window

Speech length (s)	BIC	GLR	KL2	d-vector
0.10	-	49.39%	48.45%	19.61%
0.50	38.51%	39.52%	44.18%	10.44%
1.00	26.86%	27.47%	38.78%	8.16%
1.50	20.00%	21.02%	36.47%	6.94%
2.00	15.71%	15.92%	34.74%	5.00%

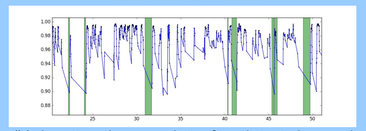
Performances in EER for different distance measures

2. Comparative score curves

- D-vector segmentation is more precise than BIC segmentation.
- With the d-vector segmentation, the distance scores change more significantly when a real speaker change occurs, it is more beneficial for peaks detection and choosing a suitable threshold to detect real speaker change point.



(a) BIC: change points often detected around the local maximum values

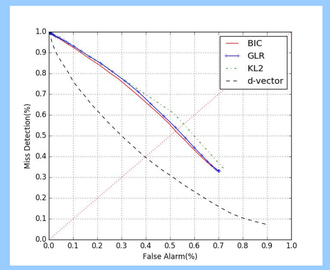


(b) d-vector: change points often detected around the local minimum values

Window distance score curves of d-vector and BIC segmentation (the green segment represents real speaker change segment)

3. Comparative segmentation evaluation in fisher dataset

- Evaluation on 100 conversations (totally 16-hours, fast speaker change scenarios, 0.3 seconds segmentation error tolerance).
- To a certain defree, d-vector is superior to the traditional segmentation approach.



DET curves of three traditional methods and d-vector based segmentation

4 Conclusion

- A noval speaker segmentation framework based on deep speaker vector.
- Can deal with the problem in fast speaker change scenarios.
- With the frame-level d-vector approach, even 0.1 seconds (10 frames) length of voice has a certain degree of speaker-discriminative ability.
- Our appraoch get more than 26% decrease in false alarm rate (FAR) and more than 21% decrease in miss detection rate (MDR) compared with traditional segmentation methods.

5 References

1. S. Chen, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in Proc. Darpa Broadcast News Transcription and Understanding Workshop, 2000, pp. 127–132
2. H.Gish, M.H.Siu, and R.Rohlicek, “Segregation of speakers for speech recognition and speaker identification,” in ICASSP, International Conference, 1991, pp. 873–876.
3. Matthew Siegler, Uday Jain, Bhiksha Raj, Stern, and Richard, “Automatic segmentation, classification and clustering of broadcastnewsaudio,” ProcDarpaSpeechRecognitionWorkshop, pp. 97–99, 1997.
4. Ke Chen and A.Salman, “Learning speaker-specific characteristics with a deep neural architecture,” IEEE Transactions on Neural Networks, vol. 22, no. 11, pp. 1744–1756, 2011.
5. Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4052–4056.

Funding Source:
National Natural Science Foundation of China under Grant No. 61271389 and No. 61371136, and the National Basic Research Program (973Program) of China under Grant No. 2013CB329302.