Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios

Lantian Li, Mingxing Xu, Thoms Fang Zheng
Center for Speech and Language Technology, Tsinghua University, China
School of Linguistic Science, Jiangsu Normal University, China

Motivation

- Speaker segmentation is difficult and many existing 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 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 difficult 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.

Methods

1. Deep speaker vectors
   - Detectchange points.
   - d-vector to the nearest seduction points in the codebook layers.
   - The DNN structure used for learning speaker-discriminative features.

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 peak detection and choosing a suitable threshold to detect real speaker change point.

3. 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 d-vectors and generate score curves.
   - Change points often detected around the local minimum values.

4. Results

   1. Speaker discriminative ability in short time window
      | Speech length (s) | BIC | CLR | ALE | DET | CER |
      |-----------------|-----|-----|-----|-----|-----|
      | 0.10            | 49.39% | 48.40% | 19.81% |
      | 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% |

   2. Comparative score curves
      - BIC: change points often detected around the local maximum values.
      - d-vector: change points often detected around the local minimum values.

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

5. Conclusion

- A novel 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 approach 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.

References


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. 2011CB3029302.