17 Who spoke when & what? How many people were there?

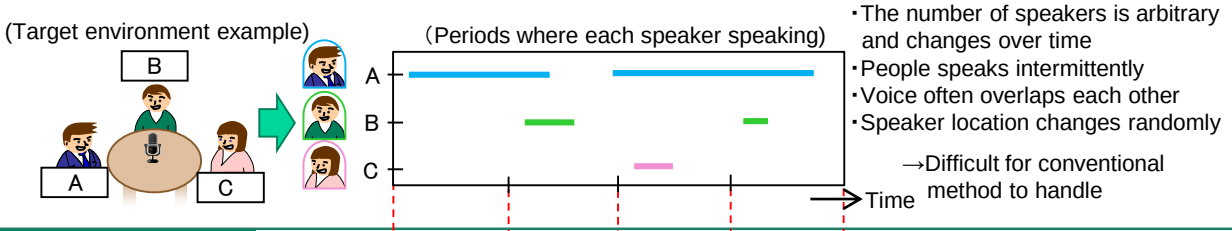
- All-neural source separation, counting and diarization model -

Abstract

We propose a method to accurately estimate "who spoke when" based on speaker's voice characteristics. It works even in a situation where multiple speaker's speech signals overlap, and accurately counts the number of speakers in such cases. Conventional methods with the similar functionality works only when the observed signal satisfies certain a priori (unrealistic) assumptions (e.g. the number of speaker known in advance, speakers never change their locations). However, these assumptions cannot be often satisfied in realistic scenarios, which leads to performance degradation. On the other hand, the proposed method, which is based purely on deep learning, can theoretically learn and deal with any realistic conversation situations. It is expected to serve as a fundamental technology for automatic conversation analysis systems, and will contribute to realization of automatic meeting minutes generation systems and communication robots.

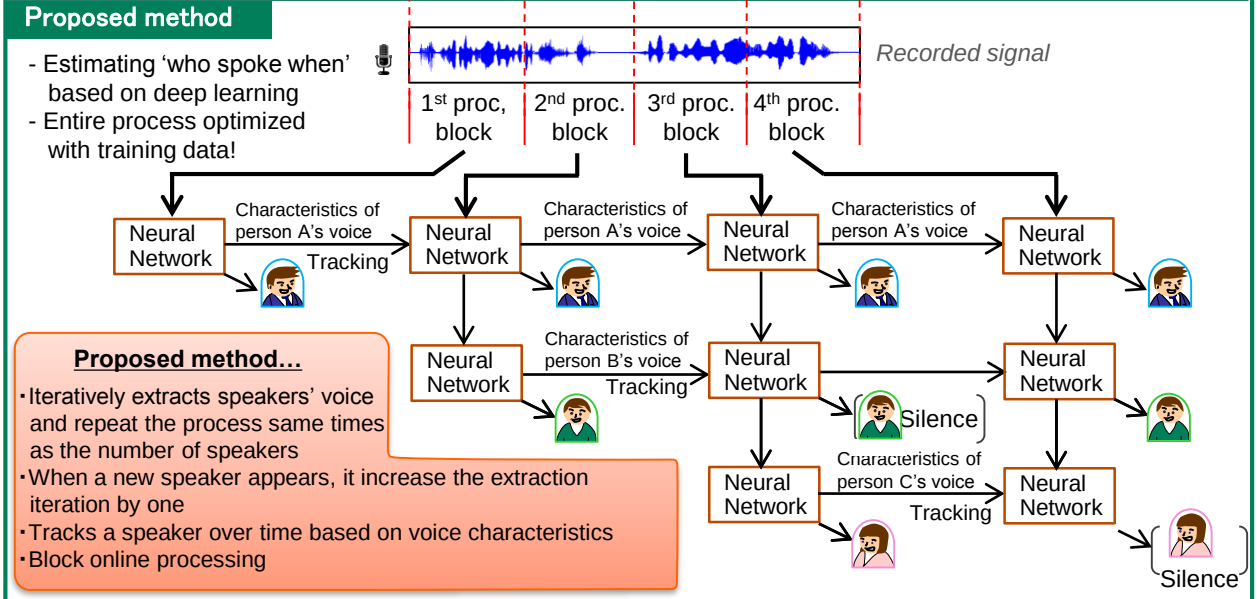
Difficulty of automatically analyzing 'who spoke when'

Conversation data is dynamic & diverse



Proposed method

- Estimating 'who spoke when' based on deep learning
- Entire process optimized with training data!



Proposed method…

・Iteratively extracts speakers' voice and repeat the process same times as the number of speakers
・When a new speaker appears, it increase the extraction iteration by one
・Tracks a speaker over time based on voice characteristics
・Block online processing

Advantage of the proposed method in comparison with conventional method

・The proposed method achieves source separation and source number counting simultaneously.
・The proposed method can track speaker's voice over time based on voice characteristics. It can keep tracking the speaker even if the speaker changes his/her location.

References

[1] K. Kinoshita, L. Drude, M. Delcroix, T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pp. 5064-5068, 2018.

[2] T. von Neuman, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, 2019.

Contact

Keisuke Kinoshita Email: cs-liaison-ml at hco.ntt.co.jp Signal Processing Research group, Media Information Laboratory

Innovative R&D by NTT

Open House 2019

Copyright © 2019 NTT. All Rights Reserved.