Université de Lorraine - IDMC
Masters SC & TAL – M1
2019-2020

Fiche de projet tutoré / Project form

Titre du projet / Title of the project

Acoustic scene classification for speaker diarization

Encadrement / Supervisors
1. équipe, laboratoire / team, lab
MULTISPEECH

1. encadrant·e principal·e (nom, email) / main supervisor (name, email)
Md Sahidullah, md.sahidullah@inria.fr

2. autres encadrant·es / other supervisors
Romain Serizel, romain.serizel@loria.fr

Description / Description

1. projet global/global project

Speech is the most convenient means for human communication. Automatic *speaker diarization* is a computer-based method of determining **"who spoke when"** in a multi-talker speech conversation. This technology has applications in many important practical problems-- for example, automatic captioning of videos, creating text transcriptions of meeting & interviews, etc.

The speaker diarization system includes different sub-systems such as *speech activity detector* (SAD), *automatic speaker verification* (ASV), *speaker clustering* (SC). The SAD module identifies the speech regions from the entire audio recording whereas ASV module computes the corresponding speaker similarity of different speech segments. Finally, SC groups different speech segments according to the speakers present in the given audio recording. The state-of-the-art speaker diarization system with deep neural network technology has shown promising results for good quality audio data. However, the performance severely degrades in the realistic conditions with noisy speech. For example, recent studies on DIHARD 2019 challenge dataset show reasonably good performance can be achieved for audio recordings from audio books, interviews in studio. But the performance severely degrades for the audio-data collected in restaurant or extracted from web videos. Speaker diarization becomes very challenging mainly due to the poor ASV performance in degraded acoustic conditions.

One possible way to improve the ASV performance is to dynamically adjust its parameters for a given acoustic condition. This can be achieved with *domain adaption* methods, however, the acoustic environment needs to be identified first. Recognizing the acoustic environment is also important for speaker clustering where the required parameters such as threshold should be ideally computed on similar kind of audio data.

With this background, the overall aim of this project is to improve the speaker diarization performance in challenging scenario by identifying the acoustic conditions of the audio recording. The students will explore different acoustic scene classification methods with an application to speaker diarization on DIHARD 2019 (single channel) dataset consisting of 11 different acoustic conditions. Since some of the acoustic conditions are more similar to each other than other conditions, the students will subsequently explore grouping of the acoustic conditions into fewer classes. This will eventually benefit the speaker recognition module with more audio data for class-dependent adaptation. Finally, the students will explore DIHARD 2019 (multi-channel) dataset where the acoustic conditions are not explicitly mentioned but can be benefitted from the recording categorization.

2. biblio. UE 705 (semestre 7)

For the semester 7, the students will get familiar with the speaker diarization and acoustic scene classification literature, dataset and state-of-the-art methods. The following papers/tech reports are suggested readings.

- Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. and Vinyals, O., 2012. Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20 (2), pp.356-370.
- Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., Liberman, M. (2019) The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines. Proc. Interspeech 2019, 978-982, DOI: 10.21437/Interspeech.2019-1268.
- Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S. and Liberman, M., 2019. Second DIHARD challenge evaluation plan. Linguistic Data Consortium, Tech. Rep.
- Baseline codes for speaker diarization on DIHARD 2019: https://github.com/iiscleap/DIHARD_2019_baseline_alltracks
- Baseline codes for acoustic scene classification: https://github.com/toni-heittola/dcase2019_task1_baseline
- Link to the papers published in DIHARD 2019 challenge: https://www.isca-speech.org/archive/Interspeech_2019/ (Locate the “The Second DIHARD Speech Diarization Challenge”)
- Link to the recently investigated acoustic scene classification methods applied to the DCASE 2019 dataset (Task A): http://dcase.community/challenge2019/task-acoustic-scene-classification-results-a#technical-reports

3. réalisation. UE 805 (semestre 8)

For the next semester, the students will experiment with baseline systems in Section 2. During this period, the evaluation of the integrated system will be made on DIHARD 2019 dataset.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

The study will be done on the second DIHARD challenge (DIHARD) 2019 dataset. MULTISPEECH team has participated in the challenge and has access to the data. The

preliminary study with the very basic acoustic scene classification techniques such as k-nearest neighbors algorithm indicate that overall speaker diarization performance can be improved by class-dependent speaker diarization. Advanced scene classification methods explored in the context of DCASE challenge are expected to play an important role for further enhancement of speaker diarization performance, specially on the realistic datasets such as DIHARD 2019.

Livrables et échéancier / Deliverable and schedule

Semester 7:

- Familiarization with state-of-the-art acoustic scene classification
- Familiarization with state-of-the-art speaker diarization
- Understanding baseline codes
- Familiarization with different data visualization tools (e.g., PCA, t-SNE, etc.)
- Presentation of the work done in Semester 7

Semester 8:

- Experiments with DIHARD dataset
- Writing project report
- Writing article for a conference
- Creating reproducible research repository

Bibliographie /References (max. 4-5)

[*il ne s'agit pas de la bibliographie complète qui sera fournie aux étudiants au début du projet mais d'une bibliographie indicative pour aider à cerner le sujet*]

1. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. and Vinyals, O., 2012. Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing, 20 (2), pp.356-370.
2. Barchiesi, D., Giannoulis, D., Stowell, D. and Plumbley, M.D., 2015. Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine, 32(3), pp.16-34.
3. Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S. and Khudanpur, S., 2018. Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. In Interspeech (pp. 2808-2812).
4. Moattar, M.H. and Homayounpour, M.M., 2012. A review on speaker diarization systems and approaches. Speech Communication, 54(10), pp.1065-1103.