Spherical Speaker Diarization System

Tuomas Kaseva, Mikko Kurimo, Aku Rouhe

January 2019

Speaker diarization refers to an automatic process which aims at answering a question "who spoke and when". The relevance of the process becomes evident when a diarization task turns manually prohibitive in terms of size of processed audio streams or/and in terms of available time to complete the task. The term "spherical" is related to our novel speaker diarization system which we are planning to release in this spring.

The system provides contributions on all prominent areas of speaker diarization research: segmentation, clustering and speaker modelling. In segmentation, we introduce a new concept called Homogenity Based Segmentation (HBS) which is designed to solve a problem of detecting speaker change boundaries and segments with overlapping speech jointly. Speaker modelling is accomplished with a deep learning architecture which borrows ideas from face and language recognition and is trained with up-to-date dataset, Voxceleb2. The architecture provides spherical speaker embeddings, representations of segments of different speakers that lie on a hypersphere, that are clustered using Spherical K-means. In order to choose optimal K, we develop a modified Silhouette criterion, which we call a heuristic Silhouette criterion. In addition, we intend to share our system freely with Python code and provide examples of the system usage by the end of March.

System evaluation is based on one of the largest, if not the largest, meeting evaluation set assembled in the speaker diarization field. This evaluation set consists of 237 meetings wich are collected from AMI and ICSI meeting corpora. With the full set, we show that the system achieves diarization error rate as 3.5% and outperforms state-of-art results on AMI subset (4.8% → 3.6%). In addition, we illustrate that the effect of even oracle HBS, at least when used in tandem with our speaker embeddings, have almost neglectable effect compared to a situation where HBS is not used at all. Interestingly, this result contradicts the consensus in the speaker diarization research field that states that the overlapping speech is one the most important factors undermining speaker diarization system performance.