Task 8
Reviewing the DIHARD Challenges
Speech Technology - COM4511/6511

May 2020

1 Introduction

Speech and language technology (SLT) research is often concerned with finding solutions with real world problems. Progress in research can come through many routes, through careful and methodical foundational research, through learning about new problems, through transfer of insights from other domains. Because SLT research is motivated by solving real tasks the field has for many years adopted a specific form to advance research: so called evaluations, challenges, competitions, and more recently shared tasks. The first of these evaluations were organised in the 1980s by NIST, the US National Institute of Standards and Technology. The mode of operation invented here was to provide participants of an evaluation with common data for training of systems. After a set period evaluation data was distributed to participants for processing. Participants then submitted the outcome for assessment at a fixed deadline. Only the organiser, NIST, could compute the scores and determine the winner of a system. Because all participants obtained the same resources, but explored different methodologies a fair comparison of methods and their suitability for a specific task could be obtained. Furthermore, as evaluations were repeated year on year, real progress could be observed in consistent scientific form. Many examples of such evaluations can be found here: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation.

Since then new forms of evaluations and competitions have been designed, using different methodologies, thereby stressing factors of competition or collaboration. In most cases there are common factors that govern these scientific events:

- A well defined training set, or range of training data sets. Training may be based on open or closed setting.
- A well defined task, including well defined scoring metrics, including tools that allow scoring.
- The existence of a well annotated development set.
- The existence of an unknown evaluation set, to be released to participants at given dates for processing only. Outputs are scored by the challenge organiser.

The objective of this task is to review a challenge designed to assess speaker diarisation, the DIHARD challenges.

2 Background

Diarisation is the task of finding 'who spoke when' an a longer piece of audio recording, without any prior information about the recording. This is relevant in multi-speaker settings where the recording may

contain utterances from several different speakers, interleaved with noise, silence or other distortions. In some situations speakers may speak at the same time. Typically the speakers are unknown beforehand. For many real world recordings this is a challenging task, and often more than one microphone is used for recording the event, for improved performance in diarisation. Diarisation is a stand-alone task that can be used for a variety of purposes - and it is also often a prerequisite for downstream processing. Several review articles on diarisation exist [1, 2, 3].

While past challenges addressed diarisation in meetings and broadcast media, the recent DIHARD challenge series was devoted to noise and complex acoustics. Two challenge campaigns have been run so far. Each challenge was associated with a special session at a conference.

- DIHARD 1
  Challenge website: https://coml.lscp.ens.fr/dihard/2018/index.html
  Interspeech 2018 special session: https://interspeech2018.org/program-special-sessions.html
- DIHARD 2
  Challenge website: https://coml.lscp.ens.fr/dihard/#evaluation
  Interspeech 2019 special session: https://www.interspeech2019.org/program/schedule/#the-second-dihard-speech-diarization-challenge-dihard-ii

3 Task

The objective of this task is to create a report, no more than 4 pages. The report should address the following questions / aspects.

- What was the task or tasks to be solved by participants ? How much flexibility was given to participants ? Were DIHARD 1 and 2 identical ?
- What methods for assessment where chosen and what was the reason behind these ?
- What was the outcome of the competitions in terms of performance difference ? How many people participated, what are they key attributes of the participants relevant for solving this task.
- Identify the best performing approaches and describe their key ideas and contributions.
- Identify at least three approaches that are fundamentally different to the best approach. Try to understand the differences and briefly explain them.
- Summarise the outcome of the competitions in the form of an executive summary. What is state of the art, what could the technology be useful for , who are leading teams, is there progress ?

Find an appropriate structure for the report, for example in scientific paper format. Avoid cutting and pasting of diagrams or tables from other papers - and find your own descriptions and words. The report should be readable for a person with only a basic understanding of speech technology.

References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] A. Joshi, M. Kumar, and P. K. Das, "Speaker diarization: A review," in *2016 International Conference on Signal Processing and Communication (ICSC)*, pp. 191–196, 2016.

[3] M. Moattar and M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, 2012.