The Second DIHARD Speech Diarization Challenge

Neville Ryant *, Alejandrina Cristia †, Kenneth Church ‡, Christopher Cieri §, Jun Du ¶, Sriram Ganapathy ″, and Mark Liberman **

1Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
2Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, Paris, France
3Baidu Research, Sunnyvale, CA, USA
4University of Science and Technology of China, Hefei, China
5Electrical Engineering Department, Indian Institute of Science, Bangalore, India

November 30, 2018
1 Proposal

Speaker diarization is the task of determining “who spoke when” in a multispeaker environment, and is an essential component of many speech recognition tasks where multiple speakers are not recorded on individual well-separated channels (e.g., extended child language acquisition data, police body cam recordings, and recordings of meetings or clinical interviews). While the state-of-the-art diarization methods work remarkably well for domains such as monologue and conversational telephone speech, as was discovered during the initial DIHARD Challenge (https://coml.lscp.ens.fr/dihard/index.html), this success does not transfer to more challenging domains such as child language recordings, speech in restaurants, and web video. Indeed, while the best performing DIHARD I system achieved diarization error rates (DER; the total percentage of reference speaker time not correctly attributed to a speaker) on the order of 3-10% for dialogues recorded in relatively clean conditions, DER rose to above 30% for meeting speech and web videos and above 40% for child language recordings and restaurant speech. Moreover, when systems were not given access to gold speech segmentation, these errors rise dramatically: 10-15% for dialogue, 40-45% for meeting speech and web videos, and above 60% for child language recordings and restaurant speech.

Although diarization for these domains remains problematic, we were encouraged by the improvement that performers made during the 4 weeks the scoring server was open: median overall DER across all submitted systems (57 systems from 13 teams) fell from 34% on day 7 to 26% on day 24 for Track 1 (diarization starting from gold segmentation) and from 44% to 40% for Track 2 (diarization from scratch). When looking across all domains comprising the evaluation set, absolute reductions in median DER from 4-10% were observed. These marked improvements in systems from those teams that participated, as well as widespread interest in access to the dev and evaluation sets from other research groups in academia and industry, lead us to believe that DIHARD filled a definite need in a field where research was stuck in a rut. Therefore, we are proposing to build on the success of DIHARD I with DIHARD II, to be held from February 14th through July 1st and reported on at a special session at Interspeech 2019 in Graz, Austria.

As with DIHARD I, this will be an open challenge with no fixed training set and multiple tracks. Tracks 1 and 2 are identical to DIHARD I and cover diarization from a single distant microphone with Track 1 (mandatory) covering diarization from a reference speech segmentation and Track 2 (optional) covering diarization from scratch. The tracks share a common development and evaluation set which spans multiple domains, including monologue, map task dialogues, broadcast interviews, sociolinguistic interviews, meeting speech, speech in restaurants, clinical recordings, extended child language acquisition recordings from LENA vests, talk radio, and YouTube videos. Recordings will consist of 5-10 minute 16 kHz monochannel FLAC files with roughly 2 hours of audio from each domain. For some domains the recordings will be new, but most audio will be taken from DIHARD I, albeit with new, higher quality diarization annotation.

Additionally, in collaboration with the organizers of the CHiME challenges, we are adding a new optional track (Track 3) for teams interested in multichannel diarization. This will consist of diarization of conversational speech from dinner parties collected from multiple microphone arrays. Given the difficulty of this data for the CHiME 5 participants, we are especially interested to see how well the strategies employed by DIHARD I performers generalize to it.

The primary metric for all system comparisons will be DER, computed without any use of collars and with explicit scoring of overlapped speech regions. Following DIHARD I, we will also report a secondary metric, which is not used for ranking purposes. Most likely, this will be a segment-based metric focusing on the performance of systems for shorter segments, which most diarization systems do particularly poorly on.

Currently, our schedule calls for a formal announcement by January 1st, release of the development and evaluation sets middle of February (6 weeks prior to the Interspeech paper registration deadline), and final results submission July 1st. All submissions must be accompanied by an accepted Interspeech paper presenting the results, which will be presented at a special session at Interspeech 2019 in Graz. For the format of this session, we propose a series of oral presentations, the first of which will be given by the challenge organizers and detail the data, tracks, metrics, and overall results (to prevent repetition of these details in the papers). Depending on the number of accepted papers (DIHARD I had 7 and we anticipate upwards of double that for DIHARD II), we may require multiple sessions.
2 Summary of impacts

- Diarization is a critical enabling technology for the analysis of conversations, meetings, interviews, and extended real-life recordings such as language acquisition data (e.g., LENA recordings) and police body-cam records.

- Diarization remains a difficult problem for which the best available technology is far from adequate. Indeed, the best performing systems in the inaugural DIHARD Challenge had overall diarization error rates of 23.7% when working from the gold speech segmentation and 35.5% when working from just the audio. Much worse performance was obtained on the harder domains and particularly with conversational speech recorded in restaurants: 49.0% DER when working from gold speech segmentation and 62.3% DER when working from just the audio. Commercially available systems marketed to researchers (e.g., LENA) fair much worse.

- Despite a great deal of recent interest in this field (2017’s JSALT workshop, the initial DIHARD Challenge in 2018, a multitude papers published in the last 3 years), there has been no major, ongoing challenge to stimulate research and facilitate performance comparisons since the last NIST RT evaluation in 2009.

- DIHARD I (also an Interspeech special session) was a success with 13 teams submitting over 50 systems and 7 accepted papers. Since the conclusion, there has been widespread interest in the data generated, both from academia and industry.

- DIHARD II will build on the success of DIHARD I with a longer challenge time frame, improved annotation, additional domains, and additional tasks. Most importantly, it will add multi-channel diarization into the mix. We anticipate it will pave the way for needed improvements to single-channel and multi-channel diarization of non-trivial materials.

3 Possible contributors of papers

The following list consists of groups that are considered likely contributors. It includes groups that participated in DIHARD I as well as those that were unable to/unaware of DIHARD, but have since expressed enthusiasm.

- Academia Sinica
  - Hung-Shin Lee (hungshinlee@gmail.com)

- Apple
  - Vikramjit Mitra (vmitra@apple.com)

- Arizona State University
  - Huan Song (hsong21@asu.edu)

- ViVoLab, Aragon Institute for Engineering Research
  - Alfonso Ortega (ortega@unizar.es)

- Brno University
  - Mireia Diez Sánchez (mireia@fit.vutbr.cz)
  - Federico Landini (landini@fit.vutbr.cz)
  - Lukáš Burget (burget@fit.vutbr.cz)
• Capio
  – Kyu Han (kyuhan215@gmail.com)

• Center for Language and Speech Processing, Johns Hopkins University
  – Gregory Sell (gsell@jhu.edu)
  – David Snyder (david.ryan.snyder@gmail.com)
  – Daniel Garcia-Romero (dgromero@jhu.edu)
  – Matthew Maciejewski (mmaciej2@jhu.edu)
  – Sanjeev Khudanpur (khudanpur@jhu.edu)

• Center for Language and Speech Technologies and Applications, Technical University of Catalunya
  – Javier Hernando (javier.hernando@upc.edu)

• Center for Robust Speech Systems, University of Texas at Dallas
  – Harishchandra Dubey (Harishchandra.Dubey@utdallas.edu)
  – John Hansen (john.hansen@utdallas.edu)

• Centro de Pesquisa e Desenvolvimento em Telecomunicações
  – Diego Augusto Silva (diegoa@cpqd.com.br)

• Computer Research Institute of Montreal (CRIM)
  – Jahangir Alam (Jahangir.Alam@crim.ca)

• Duke Kunshan University
  – Ming Li (ming.li369@duke.edu)

• EMR.AI
  – Erik Edwards (erik.edwards@emr.ai)

• EURECOM
  – Jose Patino (jose.patino@eurecom.fr)
  – Hector Delgado (Hector.Delgado@eurecom.fr)

• Google
  – Quan Wang (quanw@google.com)
  – Ignacio Lopez Moreno (elnota@google.com)

• LEAP Lab, Indian Institute of Science
  – Sriram Ganapathy (sriram.iisc@gmail.com)
  – Shobhana Ganesh (shobhana224@gmail.com)

• IBM
  – Jason Pelecanos (jupeleca@us.ibm.com)
  – Petr Fousek (petrfousek@cz.ibm.com)
  – Weizhong Zhu (zhuwe@us.ibm.com)
• Institute for Infocomm Research
  – Nguyen Trung Hieu (thnguyen@i2r.a-star.edu.sg)

• Intelligent Voice
  – Gerard Chollet (gerard.chollet@telecom-paristech.fr)

• LIMSI
  – Herve Bredin (bredin@limsi.fr)

• Lincoln Labs
  – Douglas Reynolds (dar@ll.mit.edu)

• Microsoft Research
  – Dimitrios Dimitriadis (Dimitrios.Dimitriadis@microsoft.com)

• Nanyang Technological University
  – Eng-Sion Chng (aseschng@ntu.dot.edu.sg)
  – Chenglin Xu (xuchenglin@ntu.edu.sg)

• National Technical University of Athens
  – Nikolaos Ellinas (nellinas@central.ntua.gr)

• NTIS, University of Western Bohemia
  – Zbynek Zajic (zzajic@ntis.zcu.cz)

• Oya Labs
  – Ji Ho Park (jiho@oyalabs.com)

• Queensland University of Technology
  – Sridha Sridharan (s.sridharan@qut.edu.au)

• STAR-LAB, SRI International
  – Mitchel McLaren (mitch@speech.sri.com)
  – Diego Castan (diego.castan@sri.com)
  – Martin Graciarena (martin@speech.sri.com)

• SPMI Lab, Tsinghua University
  – Li Yutian (yutian-l16@mails.tsinghua.edu.cn)

• University of Augsburg/Imperial College London
  – Bjoern Schuller (bjoern.schuller@imperial.ac.uk)

• University of Edinburgh/Quorate Technology
  – Mark Sinclair (mark.sinclair@ed.ac.uk)

• University of Zaragoza
4 Biographical information for organizers

• Kenneth Church
  Kenneth Church has worked on many topics in computational linguistics including: web search, language modeling, text analysis, spelling correction, word-sense disambiguation, terminology, translation, lexicography, compression, speech (recognition, synthesis & diarization), OCR, as well as applications that go well beyond computational linguistics such as revenue assurance and virtual integration (using screen scraping and web crawling to integrate systems that traditionally don’t talk together as well as they could such as billing and customer care). He enjoys working with large corpora such as the Associated Press newswire (1 million words per week) and even larger datasets such as telephone call detail (1-10 billion records per month) and web logs. He earned his undergraduate and graduate degrees from MIT, and has worked at AT&T, Microsoft, Hopkins, IBM, and Baidu. He was the president of ACL in 2012, and SIGDAT (the group that organizes EMNLP) from 1993 until 2011. He became an AT&T Fellow in 2001.

• Christopher Cieri
  Christopher Cieri was trained as a linguist working principally on language contact and variation in phonology. More recently, he has worked at the intersection of corpus building, linguistic analysis and human language technology development and evaluation. In 1998, he became Executive Director of the Linguistic Data Consortium (LDC), where he oversees LDC operations including the creation and distribution of hundreds of databases. His recent work focuses on social dimensions of linguistic variation, corpus building for clinical applications and the science of human linguistic annotation, especially the impact of incentives and workflows.

• Alejandrina Cristia
  Alejandrina Cristia received her PhD in Linguistics from Purdue University and did post-doctoral work on neuroimaging at the Max Planck Institute for Psycholinguistics before joining the French CNRS as a Researcher in 2013. Her publications include 45 journal articles (e.g., Psychological Science; Child Development; average impact factor 2.8) and 16 conference proceeding articles (including a best short paper award from ACL 2017; h-index scholar.google.com: 18). She is the French PI for the “Analyzing Children’s Language Environments across the World” (sites.google.com/site/aclewdid) project, which is developing cross-linguistically valid annotations and automatized analysis routines for daylong audio-recordings gathered from young children.

• Jun Du
  Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From July 2009 to June 2010, he worked with iFlytek Research. From July 2010 to January 2013, he joined Microsoft Research Asia as an associate researcher. Since February 2013, he has been with USTC as an associate professor. His research interests include speech signal processing and pattern recognition. He has published more than 80 conference and journal papers with 1000+ citations in Google Scholar.
• Sriram Ganapathy
Dr. Sriram Ganapathy is an Assistant Professor at the Electrical Engineering Dept., Indian Institute of Science, Bangalore and he leads the activities of the learning and extraction of acoustic patterns (LEAP) laboratory. Before joining as a faculty member in early 2016, he spent 4 years as a Research Staff Member at the IBM T.J. Watson Research Center in Yorktown Heights, NY, USA. He obtained his PhD from the Center for Language and Speech Processing (CLSP), Johns Hopkins University, USA. Over the past 10 years, Dr. Ganapathy has published over 60 articles in leading international journals and conferences along with a number of patents. Dr. Ganapathy also won the best tutorial speaker award in Interspeech 2014 and the Pratiksha Young Investigator award in 2017. Dr. Ganapathy is a member of the International Speech Communication Association (ISCA) and a senior member of the IEEE signal processing society. His research interests are in signal processing, machine learning, deep learning and neuroscience with applications to robust speech recognition, speech enhancement and audio analytics including biometrics.

• Mark Liberman
Mark Liberman trained as a phonetician and has worked in many areas including: corpus-based phonetics; speech and language technology; the phonology and phonetics of lexical tone, and its relationship to intonation; gestural, prosodic, morphological and syntactic ways of marking focus, and their use in discourse; formal models for linguistic annotation; information retrieval and information extraction from text. He was an undergraduate at Harvard and earned a PhD at MIT, before moving on to AT&T Bell Labs from 1975 to 1990. Since 1990 he has served as the Trustee Professor of Phonetics at the University of Pennsylvania. In 1992 he helped found the Linguistic Data Consortium (LDC), whose efforts have fueled the development and advancement of human language technology (HLT), including speech and speaker recognition, machine translation, and semantic analyses. Today, the LDC is the largest developer of shared language resources in the world, distributing more than 120,000 copies of over 2,000 databases covering 91 different languages to more than 3,600 organizations in over 70 countries.

• Neville Ryant
Neville Ryant is a researcher at the Linguistic Data Consortium (LDC) at the University of Pennsylvania, where he has worked on many topics in speech recognition including: forced alignment, speech activity detection, diarization, large scale corpus linguistics, computational paralinguistics, and automated analysis of tone. He has also supported LDC’s annotation efforts through the development of new tools for named entity recognition, sentence segmentation, and comparable corpora construction for low resource languages. He did his undergraduate and graduate studies at the University of Pennsylvania, where he focused on formal semantics and the neural basis of natural language quantifiers.