# The Second DIHARD Diarization Challenge: Dataset, task, and baselines

*Neville Ryant[1], Kenneth Church[2], Christopher Cieri[1], Alejandrina Cristia[3], Jun Du[4],*
*Sriram Ganapathy[5], Mark Liberman[1]*

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

nryant@ldc.upenn.edu

## Abstract

This paper introduces the second DIHARD challenge, the second in a series of speaker diarization challenges intended to improve the robustness of diarization systems to variation in recording equipment, noise conditions, and conversational domain. The challenge comprises four tracks evaluating diarization performance under two input conditions (single channel vs. multi-channel) and two segmentation conditions (diarization from a reference speech segmentation vs. diarization from scratch). In order to prevent participants from overtuning to a particular combination of recording conditions and conversational domain, recordings are drawn from a variety of sources ranging from read audiobooks to meeting speech, to child language acquisition recordings, to dinner parties, to web video. We describe the task and metrics, challenge design, datasets, and baseline systems for speech enhancement, speech activity detection, and diarization.

**Index Terms**: speaker diarization, speaker recognition, robust ASR, noise, conversational speech, DIHARD challenge

## 1. Introduction

Speaker diarization, often referred to as "who spoke when", is the task of determining how many speakers are present in a conversation and correctly identifying all segments for each speaker. In addition to being an interesting technical challenge, it forms an important part of the pre-processing pipeline for speech-to-text and is essential for making objective measurements of turn-taking behavior. Early work in this area was driven by the NIST Rich Transcription (RT) evaluations [1], which ran between 2002 and 2009. In addition to driving substantial performance improvements, especially for meeting speech, the RT evaluations introduced the diarization error rate (DER) metric, which remains the principal evaluation metric in this area. Since the RT evaluation series ended in 2009, diarization performance has continued to improve, though the lack of a common task has resulted in fragmentation with individual research groups focusing on different datasets or domains (e.g., conversational telephone speech [2, 3, 4, 5, 6], broadcast [7, 8], or meeting [9, 10]). At best, this has made comparing performance difficult, while at worst it may have engendered overfitting to individual domains/datasets resulting in systems that do not generalize. Moreover, the majority of this work has evaluated systems using a modified version of DER in which speech within 250 ms of reference boundaries and overlapped speech are excluded from scoring. As short segments such as

backchannels and overlapping speech are both common in conversation, this may have resulted in an over-optimistic assessment of performance even within these domains[1] [11].

It is against this backdrop that the JSALT-2017 workshop [12] and DIHARD challenges[2] emerged. The DIHARD series of challenges introduce a new common task for diarization that is intended both to facilitate comparison of current and future systems through standardized data, tasks, and metrics and promote work on robust diarization systems; that is systems, that are able to accurately handle highly interactive and overlapping speech from a range of conversational domains, while being resilient to variation in recording equipment, recording environment, reverberation, ambient noise, number of speakers, and speaker demographics. As with the NIST RT evaluations, DER is adopted as the primary evaluation metric, but without use of collars or exclusion of overlapping speech. There are no constraints on training data, with participants allowed to use any combination of public/proprietary data for system development.

The initial DIHARD challenge (DIHARD I) [13] ran during the spring of 2018 and attracted registrations from 20 teams, of which 13 submitted systems. As expected, state-of-the-art systems performed poorly, with final DER on the evaluation set for the top systems ranging from 23.73% [14] when provided with reference speech activity detection (SAD) marks to 35.51% [15] when forced to perform diarization from scratch. These error rates rates are more than double the state-of-the-art for CALL-HOME [16] at the time [4, 5]. For some domains, error rates for the best systems exceeded 49% when using reference SAD and 75% when performing diarization from scratch!

The second DIHARD Challenge (DIHARD II) [17], like its predecessor, examines diarization system performance under two SAD conditions: diarization from a supplied reference SAD and diarization from scratch. As with DIHARD I, it includes a single channel input condition utilizing wideband speech sampled from 11 demanding domains, ranging from clean, nearfield recordings of read audiobooks to extremely noisy, highly interactive, farfield recordings of speech in restaurants to child language data recorded in the home using LENA vests. Unlike DIHARD I, it additionally offers a multichannel input condition requiring participants to perform diarization from farfield microphone arrays of dinner party speech drawn

---

[1]See, for instance, the release of IBM's diarization API in 2017. The feature worked well for simple cases, but when run by users on real inputs, the performance was found to be lacking, especially for overlaps, back-channels, and short turns.

[2]https://coml.lscp.ens.fr/dihard/index.html

from the CHiME-5 corpus [18]. For the first time, we also provide participants with baseline systems for speech enhancement, SAD, and diarization, as well as results obtained with these systems for all tracks.

## 2. Tracks

The challenge features two audio input conditions:

- **Single channel** – Systems are provided with a single channel of audio for each recording. Depending on the recording source, this channel may be taken from a single distant microphone, a single channel from a distant microphone array, a mix of head-mounted or array microphones, or a mix of binaural microphones.

- **Multichannel** – Each recording session contains output from one or more distant microphone arrays, each containing multiple channels. Participants are instructed to treat the arrays separately, producing one output per array. They are free to use as few or as many of the channels on each array as they wish to perform diarization.

As system performance is strongly tied to the quality of the SAD component, we also include two SAD conditions:

- **Reference SAD** – Systems are provided with a reference speech segmentation that is generated by merging speaker turns in the reference diarization.

- **System SAD** – Systems are provided with just the raw audio input for each recording session and are responsible for producing their own speech segmentation.

Together, this yields the following four evaluation tracks:

- **Track 1** – single channel audio using reference SAD
- **Track 2** – single channel audio using system SAD
- **Track 3** – multichannel audio using reference SAD
- **Track 4** – multichannel audio using system SAD

All teams are required to register for at least one of track 1 or track 3.

## 3. Performance Metrics

As in DIHARD I, the primary metric is DER [1], which is the sum of missed speech, false alarm speech, and speaker misclassification error rates. Because systems are provided with the reference speech segmentation for tracks 1 and 3, for these tracks, it exclusively measures speaker misclassification error. This is the metric used to rank systems on the leaderboard.

For each system we also compute a secondary metric, Jaccard error rate (JER), which is newly developed for DIHARD II. JER is based on the Jaccard similarity index [19, 20], a metric commonly used to evaluate the output of image segmentation systems, which is defined as the ratio between the sizes of the intersections and unions of two sets of segments. An optimal mapping between speakers in the reference diarization and speakers in the system diarization is determined and for each pair the Jaccard index of their segmentations is computed. JER is defined as 1 minus the average of these scores, expressed as a percentage. That is, it is the mean of Eq. 1 across all reference speakers $ref$, where $TOTAL$ is the duration of the union of reference and system speaker segments, $FA$ is the total system speaker time not attributed to the reference speaker, and $MISS$ is the total reference speaker time not attributed to the system speaker. It ranges from 0% in the case where each reference

Table 1: *Overview of DIHARD II datasets. For the CHiME-5 (multichannel) data, each Kinect is treated as a separate recording.*

| Input condition | Set | Duration (hours) | # Recordings |
|---|---|---|---|
| single channel | dev | 23.81 | 192 |
| | eval | 22.49 | 194 |
| multichannel | dev | 262.41 | 105 |
| | eval | 31.24 | 12 |

speaker is paired with a system speaker with an identical segmentation to 100% in the case where none of the system speakers overlap any of the reference speakers.

$$\text{JER}_{ref} = \frac{\text{FA} + \text{MISS}}{\text{TOTAL}} \qquad (1)$$

All metrics are computed using version 1.0.1 of the *dscore* tool[3] without the use of forgiveness collars and with scoring of overlapped speech.

## 4. Datasets

### 4.1. Overview

The DIHARD II development and evaluation sets draw from a diverse set of sources exhibiting wide variation in recording equipment, recording environment, ambient noise, number of speakers, and speaker demographics. The single channel input condition (tracks 1 and 2) dataset is a superset of that used in DIHARD I, though 6 hours of additional material have been added to ensure that all domains are represented in both the development and evaluation set. Additionally, two domains where the DIHARD I annotation was deemed suspect (child language and web video) have been entirely resegmented. For the multichannel input condition (tracks 3 and 4) we use the multi-party dinner recordings originally collected for and exposed during the CHiME-5 challenge [18]. The development and evaluation sets are summarized in Table 1.

The development set includes reference diarization and speech segmentation and may be used for any purpose including system development or training. As with DIHARD I, there is no training set, with participants free to train their systems on any proprietary and/or public data. Both the development and evaluation sets will be submitted for publication via LDC at the end of the evaluation.

### 4.2. Single channel data (tracks 1 and 2)

The single channel input condition development and evaluation sets consist of selections of 5-10 minute duration samples drawn from 11 conversational domains, each including approximately 2 hours of audio. The full set of domains is described below with LDC Catalog numbers where appropriate. Unless otherwise specified, all speech is English, though not necessarily by native or even fluent speakers. All audio is distributed via LDC as 16 kHz, monochannel FLAC files.

- **audiobooks** – amateur recordings of public domain English works drawn from LibriVox; care was taken to avoid overlap with LibriSpeech [21] (unpublished)

- **broadcast interview** – student produced interviews with newsmakers of the day taken from a late 1970s college

---

[3]https://github.com/nryant/dscore

radio show; recorded on open reel tapes before being digitized and contributed to LDC (unpublished)

- **child language** – day-long recordings of 6-18 month old vocalizations collected at home by University of Rochester researchers for the SEEDLingS corpus [22]

- **clinical** – interviews with 12-16 year old children intended to determine whether or not they fit the clinical diagnosis for autism; all recordings conducted at the Center for Autism Research (CAR) of the Childrens Hospital of Philadelphia (CHOP) using a mixture of cameras and ceiling mounted microphones (unpublished)

- **courtroom** – oral arguments from the 2001 term of the U.S. Supreme Court that were digitized for the OYEZ project; recordings are summed from individual table-mounted microphones, one per speaker (unpublished)

- **map task** – recordings of map tasks in which one participant, the leader, describes a route drawn on a map to the other participant, the follower, who attempts to draw the same route on a copy of the map lacking the route and optionally lacking some landmarks; audio was recorded via close-talking microphones under quiet conditions (previously released as LDC96S38)

- **meeting** – meetings with between 3 and 7 participants, each recorded with a variety of close-talking and distant microphones, from which a single, centrally located distant microphone was selected; the development set draws from the NIST Spring 2004 Rich Transcription Evaluation (LDC2007S11 and LDC2007S12) while the evaluation set draws from previously upublished recordings conducted for the DARPA Robust Omnipresent Automatic Recognition (ROAR) project at LDC in 2001

- **restaurant** – ≈1 hour sessions involving 3-6 diners recorded on a binaural microphone worn by one participant in restaurants with varying room acoustics and noise levels; inspired by the NSF Hearables Challenge and extended by LDC for DIHARD (unpublished)

- **sociolinguistic field recordings** – sociolinguistic interviews recorded under field conditions during the 1960s and 1970s; recorded under diverse locations and conditions with subjects ranging from 15 to 81 years of age and representing diverse ethnicities, backgrounds, and dialects of world English; the development set draws from SLX (LDC2003T15) and the evaluation set from DASS (LDC2012S03 & LDC2016S05)

- **sociolinguistic lab recordings** – sociolinguistic interviews recorded as part of MIXER6 (LDC2013S03) under quiet conditions in a controlled environment; sessions were recorded with a variety of close-talking and distant microphones from which a single, centrally located distant microphone was selected

- **web video** – English and Mandarin amateur videos collected from online sharing sites (e.g., YouTube and Vimeo) as part of the Video Annotation for Speech Technologies (VAST) [23] collection (mostly unpublished)

### 4.3. Multichannel data (tracks 3 and 4)

The multichannel input condition development and evaluation sets are drawn from the CHiME-5 dinner party corpus [18], a corpus of conversational speech collected during dinner parties held in real homes. The development set combines the CHiME-5 training and development sets and encompasses 45 hours of dinner parties from 18 homes. The evaluation set is identical to the CHiME-5 evaluation set and consists of 5 hours of dinner parties from 2 homes. Each party was recorded using 6 Microsoft Kinect devices (4 channel linear arrays) distributed throughout the home in such a way that the conversation was always present on each array. Due to a combination of clock drift and random frame dropping, the Kinects within each recording session exhibit massive desynchronization, both with each other and with the binaural recording devices worn by participants. For this reason, each Kinect device is treated separately with the resulting development and evaluation sets having durations of 262.4 hours and 31.2 hours respectively. All audio is distributed via the University of Sheffield as 16 kHz WAV files.

### 4.4. Processing

A limited number of recordings contained regions carrying personal identifying information (PII), which were removed prior to publication. For the *clinical* and *restaurant* domains, this was done at LDC by low-pass filtering using a 10th order Butterworth filter with a passband of 0 to 400 Hz. To avoid abrupt transitions in the resulting waveform, the effect of the filter was gradually faded in and out at the beginning and end of the regions using a ramp of 40 ms. In the case of the *sociolinguistic field recordings* domain and the CHiME-5 data, PII was removed by the original creators of the corpora. In the former case, PII was replaced by tones of matched duration, while in the latter case it was zeroed out. PII containing regions are ignored during scoring.

### 4.5. Annotation

Reference segmentation and speaker labeling was produced by annotators at LDC using a tool equipped with playback, waveform and spectrogram display. Annotators were instructed to split on pauses > 200 ms, where a pause was defined as any stretch of time during which the speaker was not producing vocalization (e.g., backchannels, filled pauses, singing, speech errors and disfluencies, infant babbling or vocalizations, laughter, coughs, breaths, lipsmacks, and humming) of any kind. Boundaries were placed within 10 ms of the true boundary, taking care not to truncate sounds at edges of words (e.g., utterance-final fricatives). Where individual close talking microphones were available for speakers, annotation was performed separately for each speaker using their individual microphone. Due to time constraints, this manual segmentation process could not be implemented for the multichannel development data; for this data, segmentation was taken from the turn boundaries established during the original CHiME-5 transcription.

An additional post-processing step was necessary for the CHiME-5 annotation to correct for the lack of synchronization between binaural recording devices and Kinects. For each Kinect, the lag between that array and the binaural recording devices was estimated at regular intervals using normalized cross-correlation. The speech boundaries etablished by annotation on the binaural devices were then corrected for each Kinect using these estimated lags.

## 5. Baseline system

### 5.1. Speech enhancement

For speech enhancement we use a densely-connected LSTM architecture [24, 25, 26] trained to predict the ideal ratio masks (IRM) [27] of speech from log-power spectra (LPS) fea-

tures. The model is trained via progressive multi-target learning [24, 28] using 400 hours of noisy speech produced by corrupting clean utterances from WSJ0 [29] and a 50 hour Chinese speech corpus from the 863 Program [30]. Utterances were corrupted using 115 noise types [24] at 3 SNR levels (-5dB, 0dB, and 5dB). The trained models as well as scripts for applying them, are distributed through GitHub[4].

### 5.2. Beamforming

For the multichannel tracks, we use weighted delay-and-sum beamforming as implemented in BeamformIt [31]. Beamforming is applied independently for each Kinect in each session using all four channels following the CHiME-5 recipe [18].

### 5.3. Speech activity detection

The baselines for tracks 2 and 4 use WebRTC's[5] SAD as implemented in the *py-webrtc* Python package[6]. Scripts for performing SAD using the same settings used to obtain the baseline results are distributed through GitHub[4].

### 5.4. Diarization

The diarization baseline is based on the previously published Kaldi [32] recipe[7] for JHU's submission to DIHARD I [14]. At a high level, the system performs diarization by dividing each recording into short overlapping segments, extracting x-vectors [33, 34], scoring with probabilistic linear discriminant analysis (PLDA) [35], and clustering using agglomerative hierarchical clustering (AHC) [36]. In contrast to the original JHU system, we omit the Variational Bayes resegmentation step [37]. The trained models are distributed through GitHub[8].

The x-vector extractor configuration is identical to that used in previous speaker recognition and diarization systems [14, 34] with two exceptions: i) 30 dimensional mel frequency cepstral coefficient (MFCC) features are used instead of mel filterbank features; ii) the embedding layer uses 512 dimensions. MFCCs are extracted every 10 ms using a 25 ms window and mean-normalized using a 3 second sliding window. For training we use a combination of VoxCeleb 1 and VoxCeleb 2 [38, 39] augmented with additive noise and reverberation according to the recipe from [33]. Segments under 4 seconds duration are discarded, resulting in a training set with 7,323 speakers. Reverberation is added by convolution with room responses from the RIR dataset [40], while additive noises are drawn from the MUSAN dataset [41]. At test time, x-vectors are extracted from 1.5 second segments with 0.75 second overlap.

Following extraction, x-vectors are pre-processed to perform domain adaptation to the DIHARD II dataset. This is done by normalizing with a global mean and whitening transform learned from the DIHARD II development set. The whitened x-vectors are then length normalized [42] and used to train a Gaussian PLDA model [35] using a subset of VoxCeleb consisting of segments of at least 3 seconds duration. Following PLDA scoring, clustering is performed using AHC with the threshold set by minimizing DER on the development data.

Table 2: *Baseline performance (measured by DER and JER) on dev and eval sets for all tracks. The Enh. column indicates whether or not speech enhancement was applied prior to SAD.*

| Track | Enh. | DER (%) | | JER (%) | |
|---|---|---|---|---|---|
| | | Dev | Eval | Dev | Eval |
| Track 1 | no | 23.70 | 25.99 | 56.20 | 59.51 |
| Track 2 | no | 46.33 | 50.12 | 69.26 | 72.1 |
| Track 2 | yes | 38.26 | 40.86 | 62.59 | 66.60 |
| Track 3 | no | 59.73 | 50.85 | 68.00 | 65.91 |
| Track 4 | no | 87.55 | 83.41 | 88.08 | 85.12 |
| Track 4 | yes | 82.49 | 77.34 | 83.6 | 80.42 |

### 5.5. Baseline results

DER and JER of the baseline system on both the development and evaluation sets for each track are presented in Table 2. The speech enhancement module is used only for tracks 2 and 4 as a pre-processing front-end for the SAD pipeline as the diarization system did not show improvements using the enhanced audio. The scores obtained by the challenge baseline are quite high, with track 1 DER roughly in line with the performance of the best DIHARD I systems [14, 15, 25] and track 2 DER 5% higher than for DIHARD I (15% without enhancement), which we suspect reflects a combination of superior SAD components in those systems and the more careful segmentation for the child language and web video domains in DIHARD II. Error rates are noticeably higher for tracks 3 and 4, reaching 50.85% and 77.34% respectively, though, again, these rates are roughly in line with those observed for the best DIHARD I systems on the two most difficult domains in that challenge: *restaurant* and *child language*.

## 6. Conclusion

The field of speaker diarization has changed drastically in the two short years we have been running this challenge. In the lead up to DIHARD I, the research community was fragmented and most research concentrated on relatively easy datasets using forgiving evaluation metrics. This both made comparison of systems difficult and led some to believe that diarization was relatively solved and uninteresting. However, we were pleased by the response to DIHARD I, both during the evaluation and after, demonstrating that there is interest in robust diarization. This renewed energy is on display in DIHARD II, which attracted 48 registered teams from 17 countries, more than doubling the number of teams registered for DIHARD I. It is also evident in the recent announcement of the Fearless Steps challenge, which includes diarization among its tasks. We hope that this year's contributions lead to marked progress toward the goal of truly robust diarization.

## 7. Acknowledgements

# 8. References

[1] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 309–322.

[2] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.

[3] W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *Proc. ICASSP*, 2016.

[4] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *Proc. ICASSP*, 2017, pp. 4930–4934.

[5] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. ICASSP*, 2018, pp. 5239–5243.

[6] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," *Proc. ICASSP*, 2019.

[7] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. Interspeech*, 2013, pp. 1477–1481.

[8] I. Viñals, A. Ortega, J. A. V. López, A. Miguel, and E. Lleida, "Domain adaptation of PLDA models in broadcast diarization by means of unsupervised speaker clustering." in *Proc. Interspeech*, 2017, pp. 2829–2833.

[9] S. H. Yella and H. Bourlard, "Improved overlap speech diarization of meeting recordings using long-term conversational features," in *Proc. ICASSP*, 2013, pp. 7746–7750.

[10] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial neural network features for speaker diarization," in *Proc. IEEE Spoken Language Technology Workshop*, 2014, pp. 402–406.

[11] R. Milner and T. Hain, "Segment-oriented evaluation of speaker diarisation performance," in *Proc. ICASSP*, 2016, pp. 5460–5464.

[12] N. Ryant, E. Bergelson, K. Church, A. Cristia, J. Du, S. Ganapathy, S. Khudanpur, D. Kowalski, M. Krishnamoorthy, R. Kulshreshta *et al.*, "Enhancement and analysis of conversational speech: JSALT 2017," in *Proc. ICASSP*, 2018, pp. 5154–5158.

[13] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," Tech. Rep., 2018. [Online]. Available: https://zenodo.org/record/1199638

[14] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe *et al.*, "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD Challenge," in *Proc. Interspeech*, 2018, pp. 2808–2812.

[15] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolıková, O. Novotnỳ, K. Veselỳ, O. Glembek, O. Plchot *et al.*, "BUT system for DIHARD Speech Diarization Challenge 2018," in *Proc. Interspeech*, 2018, pp. 2798–2802.

[16] C. Cieri, D. Miller, and K. Walker, "From Switchboard to Fisher: Telephone collection protocols, their uses and yields," in *Proc. EUROSPEECH*, 2003.

[17] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "Second DIHARD challenge evaluation plan," Tech. Rep., 2019. [Online]. Available: https://coml.lscp.ens.fr/dihard/2019/second_dihard_eval_plan_v1.1.pdf

[18] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proc. Interspeech*, 2018, pp. 1561–1565.

[19] L. Hamers *et al.*, "Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula." *Information Processing and Management*, vol. 25, no. 3, pp. 315–18, 1989.

[20] R. Real and J. M. Vargas, "The probabilistic basis of Jaccard's index of similarity," *Systematic Biology*, vol. 45, no. 3, pp. 380–385, 1996.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[22] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," 2016, doi:10.21415/T5PK6D.

[23] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proc. LREC*, 2018.

[24] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *Proc. ICASSP*, 2018, pp. 5054–5058.

[25] L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin, and C.-H. Lee, "Speaker diarization with enhancing speech for the First DIHARD Challenge," *Proc. Interspeech*, pp. 2793–2797, 2018.

[26] L. Sun, J. Du, T. Gao, Y.-D. Lu, Y. Tsao, C.-H. Lee, and N. Ryant, "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *Proc. ICASSP*, 2018, pp. 5234–5238.

[27] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[28] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Proc. HSCMA*, 2017, pp. 136–140.

[29] J. S. Garofolo *et al.*, *CSR-I (WSJ0) Complete LDC93S6A*. Philadelphia: Linguistic Data Consortium, 1993.

[30] Y. L. Qian, S. X. Lin, Y. D. Zhang, Y. Liu, H. Liu, and Q. Liu, "An introduction to corpora resources of 863 program for Chinese language processing and human-machine interaction," *Proc. ALR*, 2004.

[31] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Language Process*, vol. 15, no. 7, pp. 2011–2022, 2007.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[33] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop*, 2016, pp. 165–170.

[34] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[35] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[36] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Trans. Audio, Speech, Language Process*, vol. 16, no. 8, pp. 1590–1601, 2008.

[37] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on Bayesian HMM with eigenvoice priors," in *Proc. Odyssey*, 2018, pp. 147–154.

[38] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[39] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *Proc. Interspeech*, pp. 1086–1090, 2018.

[40] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[41] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[42] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.