Interspeech 2024
1-5 September 2024, Kos, Greece

Evaluating the Santa Barbara Corpus: Challenges of the Breadth of Conversational Spoken Language

Matthew Maciejewski[1*], *Dominik Klement*[3*], *Ruizhe Huang*[2], *Matthew Wiesner*[1], *Sanjeev Khudanpur*[1,2]

[1]HLTCOE and [2]CLSP, Johns Hopkins University, USA
[3]Speech@FIT, Brno University of Technology, Czechia

matt@mmaciejewski.com, xkleme15@stud.fit.vutbr.cz, {ruizhe,wiesner,khudanpur}@jhu.edu

Abstract

As speech technology has matured, there has been a push towards systems that can process conversational speech, reflecting the so-called "cocktail party problem," which includes not only more challenging acoustic conditions, but also necessitates solutions to new problems, such as identifying who spoke when and processing multiple concurrent streams of speech. Such problems have been approached primarily via corpora comprising business meetings and dinner parties, overlooking the broad range of conversational dynamics and speaker demographics that fall under the category of multi-talker speech. To this end, we introduce the use of the Santa Barbara Corpus of Spoken American English for evaluation of speech technology—including preparing the corpus and annotations for automatic processing, demonstrating the failure of state-of-the-art systems to withstand the heterogeneity of conditions, and highlighting the situations where standard methods struggle to perform at all.

Index Terms: conversational speech, diarization, speech recognition

1. Introduction

Solving the cocktail party problem—the ability to recognize speech when one or more conversations are taking place, often in reverberant and noisy environments—has long been considered among the ultimate goals of the recognition of speech [1]. A variety of techniques have partially enabled this ability in speech processing systems. For instance, speech enhancement [2–5] and beamforming [6–8] have been used to suppress the noise and reverberation present in far-field recordings, and speech separation [9, 10] and overlap-aware [11] or speaker-attributed automatic speech recognition (ASR) systems [12, 13] have been developed to accommodate overlapping speech.

These techniques are often evaluated on recordings of business meetings [14–17] or dinner parties [18,19]. While work on such corpora has driven progress in multi-talker speech technology, they do not reflect the full range of human interactions or acoustic environments present in naturally occurring multi-talker audio. Recently released corpora for benchmarking audio-visual diarization [20–23] include a broader range of human interactions, but lack transcription for ASR.

One existing resource that explicitly aims to capture the full range of naturally occurring spoken interaction is the Santa Barbara Corpus of Spoken American English (SBCSAE) [24], which was originally collected for linguistic analysis of everyday speech from American English speakers of all "ages, occupations, genders, and ethnic and social backgrounds." [1] To address the shortcomings in evaluation of conversational speech technology, we repurpose the Santa Barbara corpus for evaluation of multi-talker speech technology, including speaker diarization and speaker-attributed automatic speech recognition. This involved processing of transcript files, realignment of segment boundaries, detection of anonymized regions of speech, and more. In this work, we describe our efforts in detail and also demonstrate and analyze the inconsistent performance of standard systems in these heterogeneous conditions. All corpus processing scripts and results will be released[2] and integrated into the Lhotse [25] audio preparation library for ease of use.

1.1. Related Work

Recent work in diarization has focused on creating data resources that cover the breadth of natural human interactions. For instance, The DIHARD [26] data contains some interviews, e.g., from the Mixer 6 corpus [27], as well as restaurant conversations and meetings. Recent corpora such as the AVA-AVD [22], and VoxConverse [21] datasets consist of movies and celebrity interviews and likely do not reflect everyday speech. Similar to SBCSAE is the MSDWILD [20] corpus, which used targeted web-scrapping to collect videos of everyday conversations. SBCSAE, however, has nearly double the amount of speech with many (> 5) speakers per recording.

Almost all publicly available multi-talker ASR corpora consist of meeting or dinner scenarios. While the recent CHiME-7 DASR challenge [13] focused on multi-domain speaker attributed ASR systems, the domains were all "seen" in training and limited to the CHiME [18], DipCo [19], and Mixer 6 [27], corpora which cover dinner parties and two-person interviews. SBCSAE, which comes with full transcription, can be used for evaluation of multi-talker and speaker-attributed ASR systems in more heterogeneous environments. The Ego4d-AVD set [23] has some similarities, though contains only 5-minute ego-centric multimodal clips of social interactions. SBCSAE, however, contains longer ($\sim$20 minute) recordings of *all* kinds of speech, not just social interactions, recorded in stereo.

2. Corpus Overview

The Santa Barbara corpus comprises 60 recordings (24 hours) of naturally-occurring speech, recorded from 1987–1996 with stereo microphones (though 4 recordings are monaural and 4 have a silent channel), released at 22.05 kHz.

Because the corpus was constructed for linguists to study the speech of American English speakers of all backgrounds, the available transcripts are at the level of intonation units and are very detailed. Annotation includes full transcripts with

[*]Denotes equal contribution
[1]https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
[2]https://github.com/mmaciej2/sbcsae_preparation

10.21437/Interspeech.2024-2119

utterance-level time marks, speaker labels, overlap labels, vocal style markings, inhalations/exhalations, laughter and other vocalization, and some non-speech sounds. However, as the corpus was originally intended for human viewing, there are a number of formatting issues that complicate automatic processing, including inconsistent tab and space character delimiting both across and within files as well as other typographical inconsistencies. In single-speaker regions, the utterance time marks were not precise, only placing boundaries between utterances. However, the excess silence can be reliably removed by resegmenting via forced alignment with the transcripts.

SBCSAE also comes with metadata about participants' gender, age, hometown, current state, education level, years of education, occupation, and ethnicity, and also recording information such as the recording location/room, type of conversation, and conversation summary. To protect the participants' privacy, personally-identifying information, such as name or address, had been redacted via transcript anonymization and low-pass filtering of the audio prior to release. Additionally, for some speakers, some or all metadata is missing.

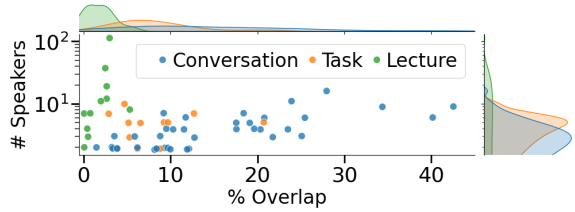

Figure 1: *Characterizing the speech in the SBCSAE by according to conversation type, the number of speakers, and the percent of overlapped speech. Each point is a recording.*

Figure 1 characterizes the styles of speech present in SBCSAE, each with unique challenges. Conversations with family, friends, romantic partners, doctors, and work colleagues feature prominently, containing between 2-10 speakers and potentially large amounts of overlapping speech. There are a number of task-oriented conversations involving game-play, cooking, veterinary clinic work, and even witness preparation. These involve groups where speech is more formal and less overlapped. Finally many recordings feature "lecture" scenarios with predominantly a single speaker (e.g. a lecturer, tour guide, or pastor), and many additional participants who speak only briefly (e.g. to ask or answer a question). Counter-intuitively, these recordings often contain the largest number of speakers.

3. Corpus Preparation

3.1. Transcript Processing and Correction

```
39.62 40.42 KENDRA:     [5(THROAT)5]
41.07 41.37                                          [6(SNIFF)6]
41.72 42.27 KEN:      [7(Hx)=7]
42.00 42.27 MARCI:    [7That wa-7] --
42.27 43.77           That made me ma=d.
43.77 45.57 KENDRA:   ... <VOX Ma=d VOX> (Hx)[=].
45.27 47.28 WENDY:                         [<VOX<SING I wa]s so= ma=d.
45.27 45.72 MARCI:                         [XXX] --
47.28 48.09           I was [2ma=d,
```

Figure 2: *Example of SBCSAE transcript file, demonstrating annotation level and style.*

An example selection from a transcript file is shown in Figure 2. Transcripts with this level of detail are valuable, but require processing to be usable for evaluation. Additionally, typographical errors can be problematic. For example, the highly-prevalent space/tab character errors lead to significant errors of tracking speakers due to difficulties in delineating the time mark entries, optionally-empty speaker attribution entry, and text transcription entry. For ASR purposes, we removed all special characters and annotations except for code-switching marks, punctuation, laughter, and undecipherable utterances. The results were then extensively checked for processing errors, and many annotation errors were corrected manually.

We made manual corrections to address inconsistent annotation: we annotated all unmarked instances of code-switching; we created—to the best of our ability—new, consistent, dummy speaker labels for speakers who were assigned a collective speaker label such as "audience". In one recording (**SBC021**), the high number of people, low volume, and short utterance length made this speaker re-annotation difficult, so we instead gave each new speaker-unlabeled utterance a new dummy label. This likely overestimates the speaker count, but better represents the dynamics of that recording. This recording should be treated carefully in diarization evaluation.

3.2. Anonymization Filter Detection

As part of the release of the Santa Barbara corpus, personal-identifying information of the participants was anonymized. This was done via text replacement in the transcripts (along with a marker) as well as filtering of speech energy above 400 Hz to maintain pitch information while removing formants necessary to recognize the words spoken. While the corpus documentation refers to "filter list files" (`*.flt`) that contain the filtered regions, we were unable to find evidence of these files having been released among the UCSB, TalkBank, and LDC copies of the corpus. As a result, we developed a manually-tuned algorithm to detect these filtered regions based on a significant reduction of energy in high-frequency parts of the spectrum. The spectra and algorithm results were manually checked and all errors found were corrected. The code and results have been released,[3] including a file in the `.uem` format used in diarization scoring packages for denoting scored regions.

3.3. Alignment

As noted in Section 2, the provided segments often include excess silence in single-speaker regions, which makes them unsuitable for evaluating speaker diarization and voice activity detection (VAD) systems; even ASR systems typically remove silence regions as a preprocessing step. We improved the ground-truth segments by re-aligning the original audio with the transcripts. Because overlapped regions are well-annotated and are challenging for forced alignment (FA) models, we only re-aligned single-speaker segments. We produced two sets of alignments: alignments designed to tightly match speech activity for diarization purposes, and ASR alignments aiming to never clip spoken words.

Our pipeline consists of three FA models: an HMM-GMM-based Montreal Forced Aligner (MFA) [28], torchaudio *-CTC-based MMS_FA [29] and WAV2VEC2_ASR_BASE_960H, both built on top of Wav2Vec2 [30]. We aligned audio on a per-segment basis for each model separately. Certain annotations corresponding to laughter, yell, or unknown words, are either deleted or mapped to "*" in the *-CTC aligners. As an orthogonal resegmentation procedure, we used the pyannote-3.1 diarization system to create a VAD-based segmentation. We computed the intersection over union (IoU) between all per-segment pairs of FA outputs and defined that two models agree if the IoU is positive. If at least one system disagreed with the others, we

[3] https://github.com/mmaciej2/sbcsae_anon_detection

left the segment for manual inspection. To obtain a per-segment matching score, we computed a weighted average of IoU scores. We assigned more weight to agreement between more dissimilar systems. If the IoU was too low or any systems do not overlap, the segments were manually inspected. Afterwards, VAD was used to confirm the presence of speech. Conflicting results were then manually inspected. The union of all alignment intervals was used as a tentative updated segment boundary. We finally applied the updated boundaries to segments that were less than half their original duration.

For diarization alignments, we aimed to remove as much silence as possible. We used the SpeechBrain CRDNN VAD system [31] in union with the pyannote diarization system, as pyannote sometimes drops large consecutive chunks of speech. We also relaxed the silence realignment condition to 10%, which in combination with VAD union allowed us to remove almost five times more silence, totalling 1 hour for the ASR and ∼5 hours for the diarization segmentations.

4. Experimental Configuration

4.1. Systems

4.1.1. Diarization Systems

We evaluated performance of state-of-the-art diarization systems on the SBCSAE using a traditional cascaded, speaker embedding clustering system and a modern end-to-end neural-based system. For the cascaded system, we used the Diarizer[4] package, based on the AMI [14] recipe. The pipeline uses pyannote-2.0's VAD and overlap detection [32–34] along with Brno University of Technology's implementation of x-vector extraction and VBx [35] with overlap [36]. We did not fine-tune any model parameters in order evaluate to what extent systems can generalize to heterogeneous conditions.

For the neural system, we used pyannote-3.1[5], for its performance record and prevalence of use both in research and production. This system is based on powerset-classification neural diarization [37], with an additional clustering component to stitch the windowed sections used to accommodate the memory requirements of neural models. For heterogeneity experiments, we used this system with all default settings. For tuning experiments, we focused on parameters which affect the number of unique speakers the model produces, evaluating both using a fixed number of speakers as well as varying the minimum number of embeddings required for a speaker cluster.

4.1.2. Speech Recognition Systems

For ASR, we sought to test robustness to challenging conditions and multi-talker overlapped speech transcription. We selected OpenAI Whisper Large-v3 [38], a single-talker ASR system pre-trained on vast amount of data. We use Guided Source Separation (GSS) [39] pre-processing to help Whisper cope with overlapped speech and enhance the audio, even though the dataset does not contain more than two channels that GSS would benefit from. As an alternative, we use the SURT 2.0 large [11] model, pre-trained on single-talker dataset LibriSpeech [40] and multi-talker simulated dataset LibriCSS [41], to test performance of an ASR system trained to transcribe overlapped speech.

We first split the long-form recordings into groups according to the re-aligned ASR supervisions (transcription units, such as parts of sentences). Then, we ensured that each group contains at most 20 re-aligned supervisions, which limits the duration of a single group to at most 1 minute as well as the maximum number of speaker changes, and restricts the MIMO WER combinatorial space.

[4]https://github.com/desh2608/diarizer

[5]https://huggingface.co/pyannote/speaker-diarization-3.1

4.2. Evaluation Metrics

4.2.1. Diarization Metrics

For evaluation, we used both Diarization Error Rate (DER) [42] and Jaccard Error Rate (JER) [26], as they have complementary downsides. The salient difference is that while DER is in some sense the amount of errors divided by the total amount of speech, JER is roughly the per-speaker detection error averaged across all reference speakers. This means that DER tends to under-emphasize errors of rare speakers due to contributing little overall speech, while JER does not penalize hallucinating extra speakers which do not correspond to a reference speaker.

4.2.2. Speech Recognition Metrics

We evaluated speaker-agnostic ASR performance using MIMO-WER [43] as implemented in the MeetEval toolkit [44], as it enables evaluating a single hypothesis stream against against multiple overlapping references. For speaker-attributed ASR we used cpWER [12].

5. Results and Discussion

5.1. Diarization Results

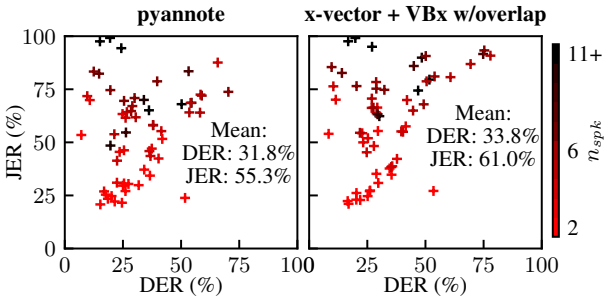

Figure 3: *DER/JER comparison on a per-sample basis.*

The baseline diarization results are reflected in Figure 3. There is a wide range of performances by both systems, with a floor of around 20%, which is not unreasonable for generally far-field recordings, and a ceiling of around 75% DER. As some recordings contain over 20 unique speakers, it is unsurprising that JER tops out above 99%. It also matches expectations that recordings with lower DER but higher JER have more speakers in them. Interestingly, there is one recording where JER is much lower than DER. The systems consistently overestimate the number of speakers in it, which aligns with the primary potential downside of JER.

It is also encouraging to see that there are recordings where both DER and JER are high, which indicates truly bad performance rather than artifacts of the metrics. The worst samples for pyannote are dominated by speech by the elderly and teenagers, likely due to a lack of age range in the training data. The x-vector system performs better on these samples, likely due to the broad range of training data in the x-vector system. However, it tends to do very poorly on heavily-overlapped recordings, suggesting the overlap-aware VBx method does not do as well as a neural model at recovering overlapping speech.

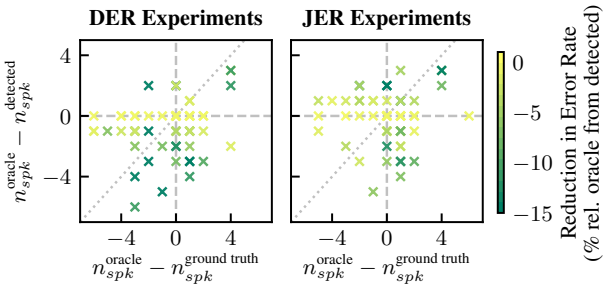

Figure 4: *Per-sample comparisons of speaker count between the automatically-detected, best-performing hard-coded, and ground truth number of speakers. The lines represent the boundaries when two of the above were equivalent.*

Additionally, we explored the speaker-counting aspect of diarization through the pyannote model, which both has default behavior of automatically detecting the number of speakers via its clustering method as well as the ability to cluster to a fixed number of speakers. We swept this parameter to find the best-performing 'oracle' number of speakers for each recording. Comparisons of the oracle, detected, and ground-truth number of speakers are shown in Figure 4.

The primary takeaway from this figure is that there is no systemic bias between these quantities—which would be seen if the points tended to lie on one side or another of the lines. One interesting result is that the *most* systemic bias is that pyannote tends to detect a larger number of speakers than is optimal for DER, but not JER. This suggests that pyannote does a reasonable job of detecting speakers, and is not optimizing the DER metric, which can incentivize dropping rare speakers.

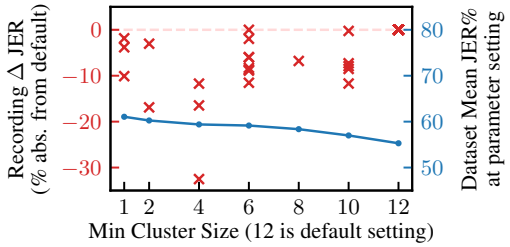

Figure 5: *Exploration of* `min_cluster_size` *parameter of pyannote diarization. Red 'x's show per-recording diarization improvement from using the 'oracle' setting, and the blue line shows the overall average error rate.*

Finally, Figure 5 reflects analysis we did with respect to the `min_cluster_size` parameter of pyannote, which helps prevent the model from over-clustering, but also provides a floor to the amount of speech the model is capable of detecting as a unique speaker. Similar to the prior experiment, we found the per-sample 'oracle' optimal setting.

In many cases, decreasing this parameter can lead to dramatic reductions in JER (up to 33% absolute), suggesting that the clustering backend does drop speakers that the end-to-end model found. However, decreasing this parameter reduces overall performance, suggesting it is already tuned optimally for this dataset (from below), even for the speaker-emphasizing JER metric. For DER, we unsurprisingly saw no gains by reducing this parameter and still saw an increased error rate in aggregate.

5.2. Speech Recognition Results

Table 1 reports the performance of Whisper and SURT Large. While the SURT model is able to detect and separate overlapped speech to an extent, it often produces incoherent sentences even on clean recordings with little overlap. This may be due to training on synthetically created mixtures from LibriSpeech [40] that do not reflect real-world conditions. We therefore performed 4-fold cross-validation, fine-tuning (FT) the SURT model on each chunk of ∼5 hours of speech. However, fine-tuning does not help, likely because SBCSAE contains such heterogeneous recordings. While Whisper was not trained to perform multispeaker ASR, and often fails to transcribe non-dominant speech in overlapped regions, it significantly outperforms the SURT model likely thanks to robustness from training on much larger amounts of data. The MIMO-WER ranges from 5.99% on mostly single speaker recordings to 42.31% on recordings with significant overlapped speech. Examining Table 1, we see that both models perform considerably better on groups with low overlap. This suggests that pre-training models on colossal amounts of single speaker data leads to superior performance, but does not solve ubiquitous problems in spontaneous speech.

Table 1: *MIMO WER% comparison using segmentations with at most 20 supervisions. The high/low overlap boundary is 10%.*

Model	Low ovl.	High ovl.	Avg.
Whisper	20.51	33.14	25.11
SURT Large	50.48	64.84	55.71
SURT Large + FT	51.62	67.52	57.38

SBCSAE also enables evaluation of speaker-attributed ASR. We first evaluated Whisper on speaker-labeled supervisions using oracle diarization, and later used GSS to assist Whisper and potentially enhance the target speaker on overlapped speech. As Figure 6 shows, GSS decreased the maximum cpWER by 11.2% absolute. The improvement mainly comes from a single recording (**SBC019**) that contains two less-correlated channels, which GSS benefits from as opposed to highly-correlated stereo recordings prevalent in the data. Furthermore, the violin plot displays a slight cpWER mass shift towards lower error rate values, demonstrating that GSS results in small improvements on other recordings as well. Finally, we evaluated Whisper + GSS with a pyannote diarization system. The last violin plot in Figure 6 shows 25.48% relative cpWER degradation when real a diarization systems is used compared to Whisper + GSS with oracle diarization. The long distribution tail shows how diarization can negatively affect cpWER, which on **SBC011** increased from 21.61% to 109.34% due to failure to distinguish between voices of the elderly.

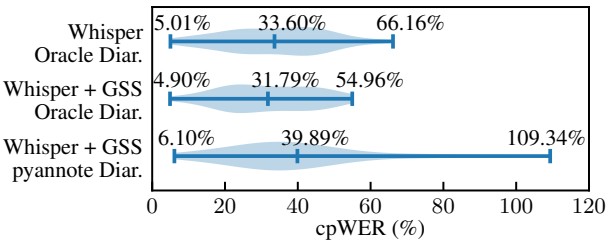

Figure 6: *Speaker-attributed ASR evaluation.*

6. Conclusion

We have presented the Santa Barbara corpus for diarization and ASR evaluation, benchmarking performance of standard speech technologies in wide conversational settings, highlighting difficulties of speaker detection in diarization and the failings of using large pre-trained models or synthetic data to tackle spontaneous speech recognition.

7. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] S. Pascual *et al.*, "SEGAN: Speech enhancement generative adversarial network," in *Proc. ISCA Interspeech*, 2017, pp. 3642–3646.

[3] H. Phan *et al.*, "Improving GANs for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[4] A. Défossez *et al.*, "Real time speech enhancement in the waveform domain," in *Proc. ISCA Interspeech*, 2020, pp. 3291–3295.

[5] B. J. Borgström and M. S. Brandstein, "Speech enhancement via attention masking network (SEAMNET): An end-to-end system for joint suppression of noise and reverberation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 515–526, 2020.

[6] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5745–5749.

[7] Z.-Q. Wang *et al.*, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *2021 IEEE Spoken Language Technology Workshop*, 2021, pp. 905–911.

[8] Z. Zhang *et al.*, "All-neural beamformer for continuous speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 6032–6036.

[9] C. Subakan *et al.*, "Attention is all you need in speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 21–25.

[10] Y. Wang *et al.*, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[11] D. Raj *et al.*, "SURT 2.0: Advances in transducer-based multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, pp. 3800–3813, 2023.

[12] S. Watanabe *et al.*, "CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[13] S. Cornell *et al.*, "The CHiME-7 DASR Challenge: Distant meeting transcription with multiple devices in diverse scenarios," in *Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023)*, 2023, pp. 1–6.

[14] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[15] A. Janin *et al.*, "The ICSI meeting corpus," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 2003.

[16] Y. Fu *et al.*, "AISHELL-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Proc. ISCA Interspeech*, 2021, pp. 3665–3669.

[17] F. Yu *et al.*, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6167–6171.

[18] J. Barker *et al.*, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ISCA Interspeech*, 2018, pp. 1561–1565.

[19] M. V. Segbroeck *et al.*, "DiPCo — dinner party corpus," in *Proc. ISCA Interspeech*, 2020, pp. 434–436.

[20] T. Liu *et al.*, "MSDWild: Multi-modal speaker diarization dataset in the wild." in *Proc. ISCA Interspeech*, 2022, pp. 1476–1480.

[21] J. S. Chung *et al.*, "Spot the conversation: Speaker diarisation in the wild," in *Proc. ISCA Interspeech*, 2020, pp. 299–303.

[22] E. Z. Xu *et al.*, "AVA-AVD: Audio-visual speaker diarization in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3838–3847.

[23] K. Grauman *et al.*, "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[24] J. W. Du Bois *et al.*, "Santa Barbara corpus of spoken American English, parts 1–4," Philadelphia: Linguistic Data Consortium, 2000–2005, Web Download.

[25] P. Żelasko *et al.*, "Lhotse: a speech data representation library for the modern deep learning ecosystem," 2021, arXiv:2110.12561.

[26] N. Ryant *et al.*, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *Proc. ISCA Interspeech*, 2019, pp. 978–982.

[27] L. Brandschain *et al.*, "Mixer-6 speech LDC2013S03," Philadelphia: Linguistic Data Consortium, 2013.

[28] M. McAuliffe *et al.*, "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proc. ISCA Interspeech*, 2017, pp. 498–502.

[29] V. Pratap *et al.*, "Scaling speech technology to 1,000+ languages," 2023, arXiv:2305.13516.

[30] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[31] M. Ravanelli *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[32] H. Bredin *et al.*, "pyannote.audio: Neural building blocks for speaker diarization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7124–7128.

[33] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. ISCA Interspeech*, 2023, pp. 1983–1987.

[34] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. ISCA Interspeech*, 2021, pp. 3111–3115.

[35] F. Landini *et al.*, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks," *Computer Speech Language*, vol. 71, p. 101254, 2022.

[36] L. Bullock *et al.*, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7114–7118.

[37] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Proc. ISCA Interspeech*, 2023, pp. 3222–3226.

[38] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," 2022, arXiv:2212.04356.

[39] D. Raj *et al.*, "GPU-accelerated guided source separation for meeting transcription," in *Proc. ISCA Interspeech*, 2023.

[40] V. Panayotov *et al.*, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[41] Z. Chen *et al.*, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7284–7288.

[42] J. G. Fiscus *et al.*, "The rich transcription 2006 spring meeting recognition evaluation," in *Machine Learning for Multimodal Interaction*, S. Renals *et al.*, Eds. Springer, 2006, pp. 309–322.

[43] T. von Neumann *et al.*, "On word error rate definitions and their efficient computation for multi-speaker speech recognition systems," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[44] ——, "MeetEval: A toolkit for computation of word error rates for meeting transcription systems," in *CHiME Workshop*, 2023.