VOXMM: RICH TRANSCRIPTION OF CONVERSATIONS IN THE WILD

Doyeop Kwak[1*], *Jaemin Jung*[1*], *Kihyun Nam*[1], *Youngjoon Jang*[1],
Jee-weon Jung[2], *Shinji Watanabe*[2], *Joon Son Chung*[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Carnegie Mellon University, USA

ABSTRACT

This paper presents a multi-modal dataset that contains rich transcriptions of spoken conversations. As diverse multi-modal and multi-task models emerge, there is a growing need for multi-modal training and evaluation datasets accompanied by rich metadata. However, there is no universal dataset that addresses these requirements for the diverse tasks partially due to the cost of annotation. To overcome this limitation, we develop a semi-automatic pipeline that makes the annotation more feasible. The resulting dataset is VoxMM, a multi-modal, multi-domain dataset. VoxMM incorporates video, audio, and text modalities. In terms of labels, it offers a wide array of metadata such as speaker labels, transcriptions, gender, and more. VoxMM supports both the training and the evaluation of any-to-any modality mapping models. It also offers a more accurate representation of real-world scenarios, bridging the gap between controlled laboratory experiments and the varying performances in the real-world. We present initial benchmarks on automatic speech recognition and speaker diarisation. The VoxMM dataset can be downloaded from https://mm.kaist.ac.kr/projects/voxmm

Index Terms— Audio-Visual, Dataset, Speech Recognition, Speaker Diarisation, Speaker Recognition

1. INTRODUCTION

Delving into the world of spoken conversations requires a depth of understanding beyond text transcription. With the advances of deep learning techniques [1–3] and the availability of large-scale datasets [4–6], Automatic Speech Recognition (ASR) technology has seen remarkable progress in recent years. While the state-of-the-art audio-only systems demonstrate impressive accuracy in controlled scenarios, a holistic understanding of real-world conversations remains a challenge.

In response to this, several multi-domain and multilingual speech datasets [7–10] have been introduced. These datasets aim to provide a more comprehensive and diverse set of data for training ASR models. Moreover, to develop models better at handling noisy conditions, recent works [11–13] have explored the potential of audio-visual speech recognition. This strand of work taps into both auditory and visual signals, significantly improving speech recognition performance in challenging conditions. This development has been accompanied by the introduction of audio-visual datasets [11, 14–16]. In particular, LRS2 [11] and LRS3-TED [16] datasets, which focus on news and lecture domains, have facilitated significant progress in speech recognition performance under noisy conditions. However, there are limitations when it comes to representing contexts, particularly concerning the speaker information.

To this end, speaker diarisation, a task that determines 'who spoke when' in multi-person conversations, has emerged as a significant research area. Researchers in this community focus on several datasets covering specific domains such as phone calls [4, 17] and meetings [18]. Although these datasets serve as valuable benchmarks for the task, the datasets do not address 'in the wild' conversations. The DIHARD speaker diarisation challenges have been introduced [19, 20] to address more challenging scenarios, offering audio data that are more representative of real-world conditions. With the growing accessibility of multimedia, additional modalities have been introduced to the task of speaker diarisation, thus moving from solely audio-based approaches to those incorporating audio-visual information. There are a number of audio-visual datasets [21, 22] that are collected in indoor environments, and the AVA-AVD [23] dataset, derived from the AVA-ActiveSpeaker dataset [24], has been introduced with the objective of representing more challenging scenarios.

For a holistic understanding of multi-talker conversations, it is crucial to have the knowledge of both what is being said, and who is saying it. While many datasets contain either text or speaker annotation, there are very few datasets that provide both (see Table 1). CHiME-6 [25] and AMI Corpus [21] offer both types of annotations, but the former only contains audio covering one specific domain, whereas the latter is audio-visual but recorded in a controlled environment.

To address these limitations, we introduce VoxMM, a multi-modal, multi-domain dataset that reflects real-world conversational scenarios. We collect video data from 12 distinct domains, encompassing a wide range of video durations ranging from 27 seconds to 10,344 seconds. This diversity in video length allows us to simulate long-term conversations. Furthermore, based on this collected data, we establish a semi-automatic annotation pipeline that significantly reduces the cost of human labeling while guaranteeing the quality of the annotation. Consequently, our dataset offers a wide array of metadata, organised with 28 distinct attributes, including but not limited to speaker labels, transcriptions, types of noise, and others. By providing this extensive metadata, our dataset has the capability to serve as a valuable resource for both training and evaluation, accommodating to various conversational tasks such as ASR, speaker diarisation, and speech enhancement. Moreover, it holds significant potential to address contemporary research trends involving multi-task scenarios. As an illustrative instance, joint ASR-diarisation task [26] demands labels that contain richer information compared to ASR and speaker diarisation tasks. In this regard, our VoxMM dataset is poised to provide substantial advantages not only for the joint ASR-diarisation task but also for numerous research directions involving multi-modal and multi-task scenarios.

Our work presents the following three contributions: (1) We design a semi-automatic audio-visual dataset creation pipeline that dramatically reduces human labor with low bias and reliable quality. (2) Using this pipeline, we have constructed an 'in the wild' audio-visual dataset of 109 hours named VoxMM, including both ASR and speaker diarisation labels. (3) Along with ASR and speaker diarisation labels, we provide

* These authors contributed equally. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT, 2022-0-00989).

Name	Modality	Domain	Ann. Method	Time	ASR	Diar
DIHARD	A	Mixed	Manual	46h		✓
CHiME-6	A	Daily Conversation	Manual	50h	✓	✓
LRS3-TED	A,V	Lecture	Semi-automatic	433h	✓	
VoxConverse	A,V	Debate, News	Semi-automatic	70h		✓
AVA-AVD	A,V	Movie	Semi-automatic	29h		✓
AMI Corpus	A,V	Meeting	Manual	100h	✓	✓
VoxMM	A,V	Mixed (12 domains)	Semi-automatic	109h	✓	✓

Table 1. Comparison to existing ASR and speaker diarisation datasets. **Ann. Method**: Annotation Method, **ASR**: Presence of ASR labels, **Diar**: Presence of speaker diarisation labels.

extensive information in the form of metadata. This rich metadata not only enhances the dataset's usefulness for the core ASR and speaker diarisation tasks but also paves the way for its application in a broader range of speech-related tasks.

2. DATASET DESCRIPTION

Statistics of VoxMM. A key feature of the VoxMM dataset is its provision of both ASR and diarisation annotations, covering a wide range of applications and research studies. Furthermore, the inclusion of 12 distinct conversation domains (*Daily conversation, Commercial, Entertainment, Interview, Movie, Lecture, Politics, News, Sports, Documentary, Presentation*, and *Remote meeting*) makes it closely aligned with real-world scenarios, setting it apart from previous literature in terms of diversity. Our dataset comprises 289 videos, with a collective duration of 109.27 hours, encompassing 76.09 hours of recorded utterances. Each video within the dataset has been carefully annotated. Notably, the diarisation labels provide information for 2,425 speakers, spanning a total of 21,796 speaker turns, while the ASR transcripts include a vocabulary size of 29,053 words. The dataset is split into *Dev* and *Test* sets, and comprehensive statistics for each of these sets are presented in Table 2.

Metadata Description. Another distinctive feature of VoxMM is the rich metadata, which enhances its adaptability and applicability across a wide spectrum of speech-related fields. Fig. 1 shows a sample of metadata. It contains four main categories, covering 28 detailed attributes (e.g. timestamp, speaker ID, transcript, face track, overlapped segments, and type of background noise). In other words, users have the flexibility to generate a customised dataset suitable to their specific needs and preferences, extending beyond ASR and speaker diarisation.

3. DATASET CREATION

3.1. Pipeline

The proposed dataset creation pipeline consists of several automated procedures designed to reduce the human labour involved in annotation. The pipeline also incorporates manual processes to further refine the label quality based on the pseudo-labels. For the manual refinement processes, 24 fluent English speakers are employed as annotators. Throughout the manual process, annotators review the videos, leveraging both visual and auditory cues to guarantee precision and efficiency.

Step 1: Video Collection. We crawl and collect a number of YouTube videos across 12 domains. During this process, we only collect videos with *Creative Commons* licenses, which grant permission for multimedia distribution. In the end, we manually select 289 videos that are closer to real-world scenarios and can cover a wide range of domains.


```
{
  "video_infos":
    {
      "category": "Entertainment",
      "age": "Recent",
      "noise": "Super Clean",
      ... ...
    }
  "speaker_info":
    {
      "id20982":
        {
          "name": "Mablean Ephriam",
          "gender": "Female",
          "utteranc_duration": 338.202,
          "on-screen": true,
          ... ...
        }
      ... ...
    }
  "statistics":
    {
      "video_duration": 1129,
      "utterance_duration": 990.2,
      "speakers": 9,
      "os_speakers": 5,
      ... ...
    }
  "segments": [
    ... ...
    {
      "segment_index": 29,
      "speaker_id": "id20986",
      "text": "{uhm} for a plane ticket (563/five hundred and sixty three)
        dollars",
      "start": 75.264,
      "end": 78.367,
      "background_noise": "Silence",
      "face":
        {
          "on-screen": true,
          "face_track_id": 25,
        }
      ... ...
    },
    ... ...
  ]
}
```


Fig. 1. A snip of the metadata file. A total of 28 attributes are organised into four categories.

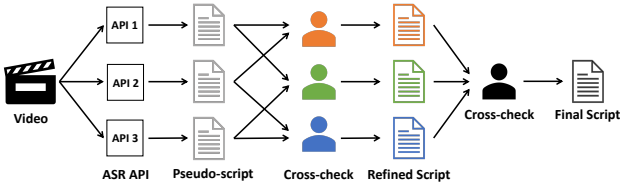
Step 2: ASR Transcription. The overall ASR transcription process is illustrated in Fig. 2(a). Three different commercial ASR APIs are used for generating three different pseudo-transcripts per video. These APIs leverage robust high-performance Voice Activity Detection (VAD) and ASR models, streamlining the automation process and ensuring relatively high-quality results.

We create three sets for the three annotators, with each set containing two unique pseudo-transcripts sourced from different APIs. Based on the cross-checking rule, scripts within each set are then compared to each other word by word, as shown in Fig. 2(b). The matching sections are merged automatically and reviewed by the annotators, whereas the non-matching sections are merged manually. For an unbiased refinement, we assign each set to a different annotator, preventing a single individual from working on multiple sets. Likewise, the three refined scripts are reviewed and merged by a fourth annotator into the final transcript. Note that the transcription of overlapping speech is not addressed during this stage. Additionally, inspired by [27], we enrich our transcript by including annotations for fillers, interjections, disfluencies, numeric notations, abbreviations, and uncertain words. These components are enclosed with special characters as **{}**: filler, interjection, and disfluency; **(/)**: numeric notation; **! !**: abbreviation; **[ ]**: uncertain words.

Step 3: Segment Reconstruction. The refined transcripts are then used for reconstructing new speech segments that can be utilised by other tasks such as speaker diarisation. To achieve this, we obtain word-level timestamps by using HuBERT XL [28] as a forced aligner, and reconstruct segments by merging the timestamps with intervals shorter than 0.25 seconds, ensuring a minimum gap of 0.25 seconds between segments.

	# videos	# mins	# IDs	video durations (s)	segment durations (s)	# spks	# os spks	speech %	overlap %	word insts.	# vocab
Dev	249	6,219	2,193	27 / 1,498 / 10,344	0.20 / 1.81 / 19.70	1 / 8.81 / 118	0 / 3.46 / 34	7.1 / 67.2 / 100	0 / 1.0 / 25.1	842,944	28,404
Test	40	337	232	53 / 506 / 2,174	0.20 / 1.64 / 14.89	2 / 5.80 / 15	1/ 3.28 / 10	19.3 / 74.4 / 93.1	0 / 3.8 / 17.5	44,235	4,862

Table 2. VoxMM dataset statistics. Entries that have 3 values are represented as min/mean/max. **# videos:** Total number of videos in the split, **# mins:** Total duration of videos in the split, **# IDs:** Total number of speakers in the split, **video durations (s):** Duration of videos, **segment durations (s):** Duration of segments, **# spks:** Number of speakers in the video, **# os speakers:** Number of on-screen speakers in the video, **speech (%):** Percentage of speech duration, **overlap (%):** Percentage of overlapping speech duration, **words insts.:** Total Number of word instances. **# vocab:** Vocabulary size.



(a) Overview of ASR transcription process.

Script 1 You know with that being said i continue my workout schedule
Script 2 You know with that being said i continued my workout schedule

(b) An example of word by word comparison, seen by each annotator.

Fig. 2. (a) Three different scripts are produced and refined concurrently, subsequently being integrated into a single, final script. Word by word comparison is performed during the cross-checking phase. (b) Matching words (black) are automatically merged whereas the non-matching words (red) are manually corrected and merged by the annotators.

A detailed distribution of segment durations can be found in Table 2.

Step 4: Diarisation Label Generation. We generate initial pseudo-labels for speaker diarisation using an off-the-shelf model [29]. This pseudo-label is manually verified and serves as an initial reference for annotating speaker ID and adjusting the timestamp. During the manual verification process, the annotators also record comprehensive details for each speaker ID, documenting additional attributes such as gender, name, profession, appearance, or any other distinctive features where it is feasible.

Step 5: Speaker ID Annotation and Timestamp Adjustment. We compare the difference between the timestamps of the reconstructed segments and the refined diarisation pseudo-labels. When the time discrepancy between them is less than 0.15 seconds and only single speaker utterances occur within that interval, the speaker ID from the pseudo-label is automatically mapped to the reconstructed segment. On the other hand, if the time difference exceeds 0.15 seconds or multiple speaker utterances are detected simultaneously, the segments are manually adjusted. Moreover, the text labels for all overlapping segments are also transcribed in this step. We then refine the timestamps for these overlapping segments based on the final transcriptions.

Step 6: Face Track Clustering. We employ Single Shot Scale-invariant Face Detector (S3FD) [30] to detect faces in each frame. These detected faces are subsequently organised into face tracks using a position-based tracker used in [22]. For each face track, seven frames are randomly selected. We then extract face embeddings from these frames using the DeepFace model [31] and compute their average. These embeddings are then clustered using Agglomerative Hierarchical Clustering [32]. We give a distance penalty for temporally overlapping face tracks since two faces appearing at the same time cannot be of the same person. Clustering is stopped if the number of clusters formed is less than twice the number of speakers in the video, or if the minimum distance surpasses 0.15.

Step 7: Face Track Mapping. Using SyncNet [33], we generate a frame-wise confidence score for each track. These scores are sorted by speaker ID and timestamp in the segment, and then aggregated and averaged for each face track. To ensure accuracy, we apply penalties to scores for short segments of less than 1 second and ignore overlapping segments. Face tracks are then attributed a speaker ID based on a confidence score with a conservative threshold to minimise false positives. Subsequently, if any face track within a cluster shares a common speaker ID with another, we assign the same ID to all face tracks within that cluster. After the speaker ID assignment is done, segments within the time frame of a face track and sharing the same speaker ID are matched with that face track.

3.2. The Challenges of 'in the wild' Data

Data Selection. The complexity of labelling presents significant obstacles to creating new 'in the wild' datasets. Previous datasets have utilised pre-existing information such as closed captions in order to alleviate annotation efforts [7, 9, 11, 15, 16]. However, such an approach inherently limits the diversity and narrows the choice in data selection, favouring the sources with rich pre-existing information. Due to this issue, we focus on collecting the diverse domains of video, regardless of pre-existing information, and initiate the transcription process from scratch.

Transcription Quality. An issue with using pseudo-labels is that the potential bias of the ASR models might be propagated to the annotators. For ASR transcriptions in particular, the challenges are not limited to pseudo-labels. As the data becomes more challenging, human transcription becomes increasingly intricate and holds the potential for individual subjectivity to interfere. As a result, we make a considerable effort to produce transcripts of high quality with minimal bias. We formulate a transcription process in which a total of three different ASR models and four individuals contribute to the transcription process for each video. This rigorous approach enables us to minimise bias and derive transcripts of exceptional quality.

Face Assignment. Many studies employ Audio-Visual Active Speaker Detection (AV-ASD) models such as SyncNet to detect speaking faces [11, 16, 22, 34, 35]. This approach is often faced with the challenge of setting an optimal confidence threshold. In previous literature, highly conservative thresholds were used to minimise false positives since it was acceptable to discard some false negatives. However, in this work, we do not discard less confident regions of the video, since we are annotating the entire video to represent a long and continuous conversation.

This requirement acts as a challenge in the construction of automation processes. In order to mitigate this issue, we move face annotation to the final step, thereby utilising the manually annotated speaker ID and timestamp labels rather than relying solely on the AV-ASD model. This strategy can offset the false negatives caused by the conservative thresholds and efficiently handle overlapping or short speech segments. Consequently, we find nearly double the number of matching faces compared to what is possible using the AV-ASD model alone.

4. EXPERIMENTS

We perform several experiments on speaker diarisation and ASR in order to demonstrate the use cases and provide baselines for VoxMM.

Model	Mode	Train set	Test set	VAD	MS ↓	FA ↓	SC ↓	DER ↓
VBx (Res101)	A	VoxCeleb1,2 CN-CELEB	AMI	Sys	11.70	1.48	2.11	15.29
				Ref	9.55	0.0	2.83	12.37
			VoxMM	Sys	3.06	13.63	4.40	21.09
				Ref	2.26	0.0	5.07	7.32
AVR-Net	AV	AVA-AVD (Single)	AVA-AVD	Sys	12.59	17.17	21.67	51.43
				Ref	2.92	0.0	22.93	25.85
			VoxMM	Sys	3.06	13.63	20.29	36.98
				Ref	2.26	0.0	18.77	21.03
		VoxCeleb1,2 AVA-AVD (Multi)	AVA-AVD	Sys	12.59	17.17	17.11	46.87
				Ref	2.92	0.0	18.18	21.11
			VoxMM	Sys	3.06	13.63	12.65	29.35
				Ref	2.26	0.0	11.18	13.43

Table 3. The results of speaker diarisation systems. VBx and AVR-Net are audio-only and audio-visual speaker diarisation systems respectively. **Sys**: *pyannote 2.0* VAD model, **Ref**: Oracle VAD derived from diarisation labels, **MS**: Missed Speech (%), **FA**: False Alarm (%), **SC**: Speaker Confusion (%), **DER**: Diarisation Error Rate (%).

Model	Mode	PT	FT	Test set	WER ↓
wav2vec 2.0	A	LL+SF+CV	LS	LS (test-clean)	3.11
				LS (test-other)	6.62
				CV	24.08
				VoxMM	28.05
AV-HuBERT	A	VoxCeleb2 + LRS3	LRS3	LRS3	2.47
				VoxMM	29.31
	AV	VoxCeleb2 + LRS3	LRS3	LRS3	1.84
				VoxMM	28.02

Table 4. The results of speech recognition models. wav2vec 2.0 and AV-HuBERT are audio-only, audio-visual speech recognition models respectively. **LL**: Libri-light, **SF**: Switchboard and Fisher, **CV**: Common Voice, **LS**: LibriSpeech, **WER**: Word Error Rate (%).

4.1. Experimental Setup

Evaluation Datasets. For speaker diarisation, we compare our dataset to AMI Corpus [21] and AVA-AVD [23]. The AMI Corpus consists of indoor meeting recordings and AVA-AVD is collected from movies, reflecting real-world scenarios. For ASR, we utilise LibriSpeech [6], Common Voice [8], and LRS3-TED (LRS3) [16]. LibriSpeech and LRS3 are ASR datasets collected from audiobooks and TED talks, respectively, while Common Voice is collected through crowdsourcing from various speakers. All experimental results for VoxMM are based on the test set.

Evaluation Metrics. To measure the accuracy of speaker diarisation, we use *dscore* tool[1] to assess the Diarisation Error Rate (DER) [36], calculated as the sum of Missed Speech (MS), False Alarm (FA), and Speaker Confusion (SC). All scores are computed with the **Fair** protocol in [37] that takes into account an overlapping speech and incorporates an acceptance margin of 0.25-second. For the evaluation of the speech recognition model's accuracy, we adopt Word Error Rate (WER).

4.2. Speaker Diarisation

To assess the acoustic conditions in our VoxMM dataset, we additionally examine the effects of voice activity detection (VAD). As reported in Table 3, we present the results using *pyannote 2.0* [38] VAD (Sys) and Oracle VAD (Ref) respectively.

Audio-only speaker diarisation. We employ the audio-only diarisation system, VBx [37], and utilise the ResNet101 model trained on VoxCeleb1 and 2 [34, 35] and CN-CELEB [39] as an x-vector extractor. When compared to the AMI corpus, VoxMM exhibits approximately a 2% higher SC (Sys: 2.29%, Ref: 2.24%) due to its broader range of speakers and diverse domains. However, when using Oracle VAD, VoxMM shows a lower DER in comparison to the AMI Corpus since VoxMM has a relatively lower overlap ratio.

Audio-visual speaker diarisation. To evaluate the audio-visual diarisation performance, we utilise Audio-Visual Relation Network (AVR-Net) [23]. We evaluate the performance of two variants of the model: one trained solely on AVA-AVD (Single) and another trained on VoxCeleb1, 2, and AVA-AVD (Multi). The difference between the two models hinges on whether they incorporate VoxCeleb1, 2, which are 'in the wild' datasets, during the training process. As shown in Table 3, the Multi model yields lower SC and DER compared to the Single model. Specifically, on the AVA-AVD test set, there is an approximate 4.6% reduction in SC (Sys: 4.56%, Ref: 4.75%), and on VoxMM, there is an approximate 7.6% reduction (Sys: 7.64%, Ref: 7.59%). This improvement can be attributed to VoxMM faithfully reflecting the real-world scenarios, much like the VoxCeleb dataset.

4.3. Speech Recognition

For ASR experiments, we remove segments that have a duration of less than 1.5 seconds or contain fewer than three words, as well as overlapping segments.

Audio-only ASR. To evaluate the test sets from various domains, we utilise robust wav2vec 2.0 [40]. We employ wav2vec 2.0 [2] LARGE model pre-trained on Libri-light (LL) [41], Switchboard [5], Fisher (SF) [4], and Common Voice (CV) and subsequently fine-tuned on LibriSpeech (LS) with Connectionist Temporal Classification (CTC) [42]. As shown in Table 4, wav2vec 2.0 demonstrates a low WER on LS test-clean and test-other, which contains relatively low-noise audio. However, when applied to CV which includes a variety of voices and accents, the wav2vec 2.0 exhibits considerably higher WER. The result of VoxMM displays a 3.97% higher WER compared to CV, which is likely attributed to its inclusion of not only various voices and accents but also a wide range of background noises. Note that we do not utilise any external language model during inference in all experiments.

Audio-visual ASR. We use AV-HuBERT [13] Base model which is pre-trained on LRS3 and VoxCeleb2, and fine-tuned on 433 hours of LRS3. As reported in Table 4, when using visual information in conjunction with audio in AV-HuBERT model, there is a 1.29% reduction in WER, indicating the role of visual information in our dataset. However, since the speaker's lip movements are visible only in certain parts of the test set and as the AV-HuBERT model is not designed to consider modality-missing scenarios, the improvement in WER is modest. Further exploration is needed in relation to missing modality research.

5. CONCLUSION

We introduce VoxMM, a multi-modal conversational dataset with rich transcription, containing 109 hours of video collected across 12 categories. Using the proposed semi-automatic annotation pipeline, we have substantially reduced human labeling costs while maintaining high annotation quality. With this, we provide baselines for ASR and speaker diarisation. The comprehensive metadata provided by VoxMM facilitates a broad range of speech-related research beyond ASR and speaker diarisation.

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

6. REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proc. NeurIPS*, 2017.

[2] A. Baevski, Y. Zhou, A. Mohamed and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[3] A. Gulati, J. Qin, C.-C. Chiu et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020.

[4] C. Cieri, D. Graff, O. Kimball et al., "Fisher english training speech parts 1 and 2 ldc200," *Linguistic Data Consortium*, vol. 2005, 2004.

[5] J. Godfrey and E. Holliman, "Switchboard-1 release 2 ldc97s62," *Linguistic Data Consortium*, p. 34, 1993.

[6] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. Interspeech*, 2015.

[7] G. Chen, S. Chai, G. Wang et al., "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Proc. Interspeech*, 2021.

[8] R. Ardila, M. Branson, K. Davis et al., "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2019.

[9] D. Galvez, G. Diamos, J. Ciro et al., "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," in *Proc. NeurIPS*, 2021.

[10] C. Wang, M. Riviere, A. Lee et al., "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL*, 2021.

[11] T. Afouras, J. S. Chung, A. Senior et al., "Deep audio-visual speech recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[12] P. Ma, S. Petridis and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *Proc. ICASSP*, 2021.

[13] B. Shi, W.-N. Hsu and A. Mohamed, "Robust self-supervised audio-visual speech recognition," in *Proc. Interspeech*, 2022.

[14] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Trans. on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[15] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, 2017.

[16] T. Afouras, J. S. Chung and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv:1809.00496*, 2018.

[17] A. Canavan, D. Graff and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.

[18] A. Janin, D. Baron, J. Edwards et al., "The icsi meeting corpus," in *Proc. ICASSP*, 2003.

[19] N. Ryant, K. Church, C. Cieri et al., "The second dihard diarization challenge: Dataset, task, and baselines," in *Proc. Interspeech*, 2019.

[20] N. Ryant, P. Singh, V. Krishnamohan et al., "The third dihard diarization challenge," *arXiv:2012.01477*, 2020.

[21] J. Carletta, S. Ashby, S. Bourban et al., "The ami meeting corpus: A pre-announcement," in *Proc. MLMI*, 2005.

[22] J. S. Chung, J. Huh, A. Nagrani et al., "Spot the conversation: speaker diarisation in the wild," in *Proc. Interspeech*, 2020.

[23] E. Z. Xu, Z. Song, S. Tsutsui et al., "Ava-avd: Audio-visual speaker diarization in the wild," in *Proc. ACM MM*, 2022.

[24] J. Roth, S. Chaudhuri, O. Klejch et al., "Ava active speaker: An audio-visual dataset for active speaker detection," in *Proc. ICASSP*, 2020.

[25] S. Watanabe, M. Mandel, J. Barker et al., "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. The 6th International Workshop on Speech Processing in Everyday Environments*, 2020.

[26] L. E. Shafey, H. Soltau and I. Shafran, "Joint speech recognition and speaker diarization via sequence transduction," in *Proc. Interspeech*, 2019.

[27] J.-U. Bang, S. Yun, S.-H. Kim et al., "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition," *Applied Sciences*, vol. 10, no. 19, pp. 6936, 2020.

[28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[29] Y. Kwon, J.-w. Jung, H.-S. Heo et al., "Adapting speaker embeddings for speaker diarisation," in *Proc. Interspeech*, 2021.

[30] S. Zhang, X. Zhu, Z. Lei et al., "S3fd: Single shot scale-invariant face detector," in *Proc. ICCV*, 2017.

[31] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *Proc. ASYU*, 2020.

[32] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of classification*, vol. 1, pp. 7–24, 1984.

[33] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2017.

[34] A. Nagrani, J. S. Chung and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017.

[35] J. S. Chung, A. Nagrani and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[36] J. G. Fiscus, J. Ajot, M. Michel and J. S. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. MLMI*, 2006.

[37] F. Landini, J. Profant, M. Diez and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, pp. 101254, 2022.

[38] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, 2021.

[39] Y. Fan, J. Kang, L. Li et al., "Cn-celeb: a challenging chinese speaker recognition dataset," in *Proc. ICASSP*, 2020.

[40] W.-N. Hsu, A. Sriram, A. Baevski et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech*, 2021.

[41] J. Kahn, M. Rivière, W. Zheng et al., "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. Interspeech*, 2020.

[42] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.