# Spot the conversation: speaker diarisation in the wild

*Joon Son Chung*[1,2]*, *Jaesung Huh*[1,2]*, *Arsha Nagrani*[1]*, *Triantafyllos Afouras*[1], *Andrew Zisserman*[1]

[1]Visual Geometry Group, Department of Engineering Science, University of Oxford, UK
[2]Naver Corporation, South Korea

http://www.robots.ox.ac.uk/~vgg/data/voxconverse

## Abstract

The goal of this paper is speaker diarisation of videos collected 'in the wild'.

We make three key contributions. First, we propose an automatic audio-visual diarisation method for YouTube videos. Our method consists of active speaker detection using audio-visual methods and speaker verification using self-enrolled speaker models. Second, we integrate our method into a semi-automatic dataset creation pipeline which significantly reduces the number of hours required to annotate videos with diarisation labels. Finally, we use this pipeline to create a large-scale diarisation dataset called VoxConverse, collected from 'in the wild' videos, which we will release publicly to the research community. Our dataset consists of overlapping speech, a large and diverse speaker pool, and challenging background conditions.

**Index Terms**: speaker diarisation, speaker recognition.

## 1. Introduction

Speaker diarisation is the challenging task of breaking up multi-speaker video into homogeneous single speaker segments, effectively solving *"who spoke when"*. Beyond being an interesting research problem in itself, it is also a valuable pre-processing step for a number of applications, including speech-to-text.

While state-of-the-art diarisation systems perform remarkably well for speech from constrained domains (e.g. conversational telephone speech [1, 2, 3, 4] or meeting speech [5]), this success does not transfer to more challenging conditions found in online videos 'in the wild'. The challenges here include the lack of a fixed domain (videos can be from talk shows, news broadcasts, celebrity interviews, home vlogs), a large number of speakers (some of whom are off-screen), short rapid exchanges with cross-talk, and background degradation consisting of channel noise, laughter and applause.

These conditions make manual annotation of online videos a daunting task for human annotators, leading to a dearth of large-scale public diarisation datasets of unconstrained speech. While large-scale evaluations are held regularly by the National Institute of Standards in Technology (NIST-RTE), these are limited to constrained audio-only datasets, which are not freely available to the research community (Table 1).

To attempt to remedy some these issues, the DIHARD challenges [6, 7] were introduced in 2018. These are valuable annual challenges that cover 11 different data domains, including mother-child conversations, meetings and courtroom settings. One of these domains is also web videos, however there is a limited amount of data (only 2 hours). The datasets are also audio-only, and are only available to challenge participants (not released freely to the research community).

A large-scale diarisation dataset of videos 'in the wild' would encourage the development of new audio-visual diarisation techniques that deal with unconstrained conditions. Inspired by the recent success of automatic audio-visual dataset creation pipelines (VoxCeleb [8, 9, 10], VGGSound [11]), we propose a scalable, audio-visual method for speaker diarisation in web videos. Our method relies heavily on the recent successes of active speaker detection [12] and face and speaker verification [13, 14, 15]. We then integrate this method into a semi-automatic dataset creation pipeline – consisting of both automatic annotation and manual verification. We use this pipeline to curate VoxConverse, a challenging and diverse speaker diarisation dataset from 'in the wild' videos.

Our automatic diarisation method exploits the following three key ideas; Firstly, the speech for on-screen identities can be accurately segmented automatically using active speaker detection and then identified using face recognition, the core basis for the VoxCeleb pipeline [10]. Second, there has been great progress in creating audio-visual speech enhancement models [16, 17, 18], which separate overlapping speech into single speaker streams. Given the amount of cross-talk and background noise in web videos, we use this model to better isolate and identify speaker identities. The above two ideas allow us to accurately identify and isolate speech for on-screen speakers. Finally, to accurately recognise *off-screen* speakers, we utilise state of the art speaker recognition embeddings that verify identities from audio alone (Figure 1).

Concretely, we make the following three contributions: (i) We create an automatic audio-visual diarisation method using active speaker detection, face recognition, speech enhancement and audio-only speaker recognition. (ii) We integrate our method into a semi-automatic dataset creation pipeline which consists of human annotation and automatic diarisation. Our pipeline is scalable, and significantly reduces the number of hours required to annotate videos. (iii) We use this pipeline to curate VoxConverse, a challenging 'in the wild' audio-visual diarisation dataset. We compare our audio-visual diarisation method to existing audio-only baselines on our dataset, and show that large performance gains can be obtained from integrating visual information.

## 2. Related works

Speaker diarisation has been an active field of research for many years, but remains one of the most challenging tasks in speech processing. Deep learning techniques have not been applied to speaker diarisation to the same degree that they have for other tasks, partially due to the lack of end-to-end models for diarisation, but also due to the lack of diverse, large-scale datasets like ImageNet [19] and VoxCeleb [8].

---

* These authors contributed equally to this work.

Much of the progress in the field has been driven by a series of NIST Rich Transcription challenges (NIST-RTE), which focuses almost solely on the meeting domain. The series also proposed the diarisation error rate (DER) as an evaluation metric for speaker diarisation, which is now used as the primary metric across all domains and evaluations. Research into speaker diarisation has largely evolved independently for different domains, with broadcast news [20], telephone speech [21], and meetings [22, 23] being the most popular domains. For each domain, specific datasets have been introduced and used, all created by manual annotation. We provide a summary in Table 1.

The DIHARD series of challenges [6, 7] were introduced to overcome the domain dependency in the field – the data consists of recordings from different conversational domains, including audiobooks, broadcast interviews, child speech and so on. The evaluation conditions are challenging, and even the best performing systems score relatively high diarisation error rates of around 20% with ground truth voice activity detector (VAD), and 30% with system VAD. The annotation is performed with very fine granularity, which allows evaluation without a forgiveness collar. Barring DIHARD, all other datasets and evaluations include a generous forgiveness collar and exclude overlapping speech from scoring. Inspired by DIHARD, we annotate overlapping speech in `VoxConverse` and include it in evaluation. For almost all existing datasets, annotation is done manually and solely using the audio. Annotation without visual information is challenging, particularly when the number of speakers is large, since it is easy to be confused between voices without the additional identity redundancy provided by the face. Unlike other works, our dataset creation pipeline is semi-automatic, scalable and audio-visual.

Table 1: *Comparison to existing speaker diarisation datasets.* **Cond.:** *Acoustic conditions;* **Ann. Methd:** *Annotation Method;* †*: Fisher English Training Speech part I and II.*

| Name | Cond. | Free | Ann. Methd |
|---|---|---|---|
| 2005 NIST RTE | Meetings | ✗ | Manual |
| CALLHOME [21] | Telephony | ✗ | Manual |
| AMI Meeting Corpus [23] | Meetings | ✓ | Manual |
| ICSI Meeting Corpus [22] | Meetings | ✓ | Manual |
| Fisher† I and II [24] | Telephony | ✗ | Manual |
| DIHARD [6, 7] | Mixed | ✗ | Manual |
| **VoxConverse** | Multi-media | ✓ | Semi-automatic |

## 3. Dataset description

The development set of `VoxConverse` consists of 216 multi-speaker videos covering 1,218 minutes with 8,268 speaker turns annotated. The test set contains approximately 53 hours and will be released after the VoxCeleb Speaker Recognition Challenge in October 2020. The statistics of the dataset can be seen in Table 2.

Videos included in the dataset are shot in a large number of challenging multi-speaker acoustic environments, including political debates, panel discussions, celebrity interviews, comedy news segments and talk shows. This provides a number of background degradations, including dynamic environmental noise with some speech-like characteristics, such as laughter and applause. Our dataset is audio-visual, and contains face detections and tracks as part of the annotation.

The videos in the datasets consist of quick, short speech segments. On average, 91% of the video time contains speech, and 3–4% of this contains speech where one speaker overlaps with another speaker. The overlap percentage varies between videos; one video for example has an overlap percentage of 30.1%. Videos vary in length from 22 seconds to 20 minutes. Unlike other domains such as telephony, each video has on average between 4 and 6 speakers, with one video in the dataset having 21 speakers.

## 4. Dataset collection

The dataset collection process consists of two stages – initial annotations are generated automatically using our proposed audio-visual method, and the annotations are then checked and refined by human annotators.

### 4.1. Automatic pipeline

The automatic computer vision pipeline to curate `VoxConverse` is similar to that used to compile `VoxCeleb1` [8] and `VoxCeleb2` [9].

**Stage 1. Collection of videos.** The first stage is to obtain a list of videos. We start from a number of keywords including 'panel debate' and 'discussion' in order to obtain videos where multiple people are talking alternately or at the same time. The list of videos is obtained by searching the keywords on YouTube, and duplicate URLs that appear in the search results of multiple keywords is removed. Moreover, we remove the videos that are identical or very similar in content based on tf-idf features [25] extracted from the YouTube auto-generated subtitles. The list contains a range of videos, ranging from US presidential debates and talk shows to documentaries.

**Stage 2. Shot detection.** Shot boundaries are then determined to find within-shot frames for which face tracking is to be run. The boundaries are found by comparing intensity and brightness across consecutive frames [26].

**Stage 3. Face detection and tracking.** A CNN face detector based on the Single Shot Scale-invariant Face Detector (S3FD) [27] is used to detect faces on every frame of the video. This detector allows the detection of faces at various scales and poses. Within each shot, face detections are grouped together into face tracks using a position-based tracker, as in [8, 9].

**Stage 4. Face-track clustering.** A face recognition CNN is used to extract embeddings for every face track. The network used here is based on the ResNet-50 [28] trained on the VGGFace2 dataset. The embeddings are extracted 5 times per face track at uniform intervals, and then averaged. The embeddings are clustered using Agglomerative Hierarchical Clustering [29], but a large penalty is added to the distance matrix between overlapping face track so that they are never clustered together.

**Stage 5. Active speaker detection (ASD).** The goal of this stage is to determine if the visible face is the speaker. Two systems are used for this purpose. The first method uses a variant [30] of SyncNet [31], which is a two-stream CNN that determines the active speaker by estimating the correlation between the audio track and the mouth motion of the video. The second method isolates the speech of the target speaker from a mixture of sounds using an audio-visual speech enhancement (AVSE) network [16] then uses an off-the-shelf voice activity detector, WebRTC [32], to determine the speech segment.

Each method has its weaknesses – the SyncNet ASD acti-

Table 2: *VoxConverse dataset statistics. Entries that have 3 values are reported as min/mean/max.* **# spks:** *Number of unique speakers per video.* **# mins:** *Total number of minutes in the dataset.* **video durations (s):** *Length of videos in seconds.* **speech %:** *Percentage of video time that is speech.* **overlap %:** *Percentage of speech per video when 2 or more speakers overlap.*

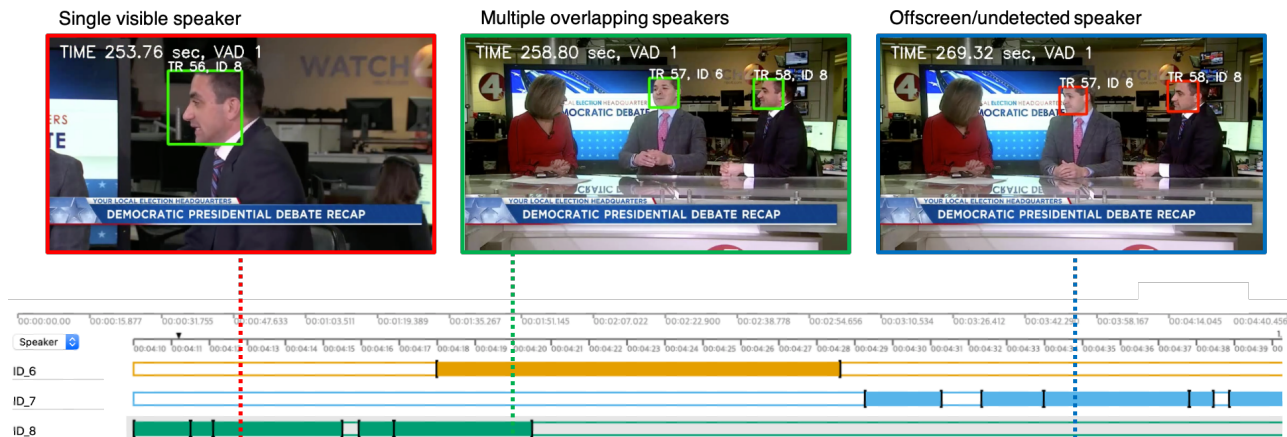| set | # videos | # mins | # spks | video durations (s) | speech % | overlap % |
|-----|----------|--------|--------|---------------------|----------|-----------|
| Dev | 216 | 1,218 | 1 / 4.5 / 20 | 22.0 / 338.2 / 1097.4 | 10.7 / 93.2 / 99.8 | 0 / 3.8 / 28.7 |
| Test | 310 | 3,210 | 1 / 6.3 / 21 | 26.0 / 621.2 / 1200.0 | 11.5 / 89.8 / 100 | 0 / 3.0 / 30.1 |



Figure 1: *Output of our automatic audio-visual diarisation method. Green squares on the images represent face detections with positive ASD output, red squares represent face detections with negative ASD output. The identities are labelled as ID_6, ID_7 and ID_8, and speaker timelines show when each identity is speaking. For clarity, we only show 3 frames from the video. Our method elegantly deals with visual speakers, overlapping speech and undetected/off-screen speakers.*

vates when the phoneme in the background speech matches the viseme shown on the target face since the model does not consider temporal context; the WebRTC voice activity detector is often activated from the residual signal left in the AVSE output which, despite the reduced power, causes false alarms. Therefore, a face track is considered to be speaking only if both of the methods agree, which helps reduce false alarms from laughter and music.

**Stage 6. Labelling off-screen speech.** A pre-trained speaker recognition model [33] is used to verify the identity of speech that comes from off-screen speakers. Any parts of the audio with positive voice activity detector (VAD) output, but with no visible active speaker is considered to be an off-screen speech segment. Speaker embeddings are extracted for the whole video, then the off-screen speech segments are compared to all speech segments with visible active speaker using the cosine distance between the embeddings. If the cosine distance is below a threshold, the off-screen segment is assigned to the speaker; if not, the segment is left as unknown for the human annotator to verify. This procedure is closely related to the multimodal diarisation method of [34].

**Discussion.** In creating the VoxCeleb datasets, very conservative thresholds were used in both active speaker detection and face verification, since it was necessary to be very certain about the speaker labels without any human intervention. This is, however, at the cost of high false rejection, which meant that a large number of true speech segments were discarded.

In contrast, a speaker diarisation dataset must contain continuous audio recording with different identities speaking in turns. Therefore, we cannot discard parts of video based on

low confidence, but the entire video must be labelled in full. The thresholds are optimised to minimise the overall Diarisation Error Rate (DER) (see section 5), since high false alarms and high false rejections both lead to increased man-hours during the manual verification and correction stage. Note how our pipeline consists of two ASD methods, as we show later, this redundancy is beneficial for performance.

### 4.2. Manual verification

The output of the automatic pipeline has been checked and corrected by the authors of this paper using a customised version of the VGG Image Annotator [35, 36]. This was done so that the authors can identify the failure modes and make guidelines for the external annotators when the process is scaled up. The tool allows the user to watch and verify the annotation at various speeds, and with aid of video.

During the annotation process, a number of failure modes were identified. The most common is non-visible speech segment assigned to the wrong speaker, but false alarm of the VAD and missed overlapped speech are also relatively common.

**Guidelines.** Speech segments are split when pauses are greater than 0.25 seconds. Unlike some previous datasets in diarisation, laughter is not assigned to identities, as it is difficult to assign an accurate label to audience laughter. Anything that can be transcribed, including short utterances such as 'yes' and 'right', are considered to be speech. Known speakers are named in the annotation process to facilitate easier cross-checking. The annotators are asked to be as careful as possible that the marked boundaries are within 0.1 seconds of the true boundary.

**Quality check.** In order to verify the quality of manually

checked annotations, a subset of the data has been labelled independently by two different annotators. This subset contains 1 hour of material from 15 YouTube videos. The diarisation error rate between the two annotations is approximately 1%, using the labels from one annotator as the reference and the other as the prediction. This error can be mostly attributed to disagreements on the source of off-screen speech segments.

**Discussion.** Diarisation labels for 'in the wild' conversations are difficult to obtain. It is almost impossible to manually annotate the segments in our dataset without the video. Even with the video, it can take 10 times the video duration to annotate segments to satisfactory quality if *starting from scratch*, particularly for many speakers. In contrast, the verification of our audio-visual method output takes around twice the video duration, and is possible with less experienced annotators.

The time taken to annotate correlates strongly to the quality of the output from the automatic method. The first few videos in the development set were annotated with initial hyperparameters that gave relatively poor performance. The diarisation labels were then manually fixed, and the parameters were re-tuned on this data to minimise the diarisation error rate. More videos were then generated using the new set of parameters and this process was repeated a few times. While it is possible that some types of errors are more time-consuming for humans to fix compared to others, we have observed that the annotation became faster after each iteration.

## 5. Experiments

We compare our audio-visual method to an audio-only DIHARD 2019 baseline, and also compare performance to two ablations.

**DIHARD 2019 baseline.** The second DIHARD [6, 7] challenge provides a baseline system based on the JHU submission of the first DIHARD challenge. We use this public code[1] as an audio-only baseline.

The overall procedure is as follows. Speech segments are obtained using VAD, and divided into short overlapping segments (1.5s with 0.75s overlap). Speaker embeddings are extracted using the x-vector [37] system, and the similarities between the embeddings are scored with a pre-trained probabilistic linear discriminant analysis (PLDA) [38, 39] model also provided in the code. Segments are then grouped using agglomerative hierarchical clustering (AHC) based on PLDA scores. We report the best performance by tuning the threshold of the AHC on the development set.

Two variants are compared, with and without the speech enhancement module [40] which has been made publicly available[2]. The system uses a Long short-term memory (LSTM) based speech denoising model trained on simulated training data. This model shows state-of-the-art performance on speech enhancement, and has shown its effectiveness for diarisation in the first DIHARD challenge.

**Ablations.** A crucial design choice that we made is that we used two active speaker detection methods, and a segment was only marked positive when both methods gave a positive output. We consider two ablations of our method – one using only SyncNet-based ASD, and the other using only AVSE-based ASD.

**Evaluation protocol.** Methods are evaluated on the `VoxConverse` development set. We use the diarisation er-

Table 3: *Results on the dev set using baseline methods and our proposed audio-visual method. All values are in %.* **MS:** *missed speech;* **FA:** *false alarm;* **SC:** *speaker confusion;* **DER:** *diarsation error rate (where $DER = MS + FA + SC$). For each metric, the lower the better.* † *Audio-only baselines.*

| Name | MS | FA | SC | DER |
|------|----|----|----|-----|
| DIHARD 2019 baseline [6] † | 11.1 | 1.4 | 11.3 | 23.8 |
| DIHARD 2019 baseline w/ SE [6, 40] † | 9.3 | 1.3 | 9.7 | 20.2 |
| Ours (SyncNet ASD only) | 2.2 | 4.1 | 4.0 | 10.4 |
| Ours (AVSE ASD only) | 2.0 | 5.9 | 4.6 | 12.4 |
| **Ours (proposed)** | 2.4 | 2.3 | 3.0 | 7.7 |

ror rate (DER), defined as the sum of missed speech (MS), false alarm speech (FA), and speaker misclassification error (speaker confusion, SC). A forgiveness collar of 0.25 seconds is applied in order to compensate for small inconsistencies in annotation.

**Training.** All thresholds are tuned on the `VoxConverse` development set. The AHC threshold for speaker clustering is the only hyperparameter to be tuned in the audio-only baseline. The audio-visual method requires three key thresholds – cosine distance for face clustering, SyncNet confidence for active speaker detection, and cosine distance for speaker identification. The first of these affect performance the most, since any error in the identity clustering directly causes speaker confusion.

**Results.** Table 3 shows the results of all the evaluations. Our audio-visual method obtains a DER much lower than the audio-only state-of-the-art baselines, showing the efficacy of using visual information for diarisation on this dataset. The ablation analysis for the ASD methods proves the effectiveness of using two active speaker detectors – the combined method has a significant decrease in false alarm rate for only a small increase in missed speech.

With regards to the difficulty of `VoxConverse`, we note that the DIHARD 2019 baseline obtains a DER of about 20% on our dataset (Table 3), and hence there is a lot of room for improvement. While this is lower than the 26% that the same model achieves on the extremely challenging DIHARD development set (with ground truth VAD), we hypothesize that this difference may be attributed to the use of a 0.25-second forgiveness collar in our evaluation protocol.

## 6. Conclusion

We have developed a high performance audio-visual algorithm for automated diarisation, and used it to generate a new speaker diarisation dataset, `VoxConverse`, from 'in the wild' videos. The pipeline is fully scalable and effective across a range of domains. `VoxConverse` currently contains 70 hours of annotated video, but we are in the process of scaling up. The data will be used in the second VoxCeleb Speaker Recognition Challenge in October 2020 and, after that, will be released publicly to the research community free of charge.

## 7. Acknowledgements.

# 8. References

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

[2] W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *Proc. ICASSP*. IEEE, 2016, pp. 5045–5049.

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

[4] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *Proc. ICASSP*. IEEE, 2019, pp. 6301–6305.

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

[6] 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 *Interspeech*, vol. 2018, 2018, pp. 2808–2812.

[7] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," in *Interspeech*, 2019.

[8] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[9] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[10] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[11] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: a large-scale audio-visual dataset," in *Proc. ICASSP*. IEEE, 2020.

[12] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC.*, 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015.

[15] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. ICASSP*, 2019.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *INTERSPEECH*, 2018.

[17] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[18] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Interspeech*, 2019.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[20] S. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," in *Speaker Odyssey*, 2004.

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

[22] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *Proc. ICASSP*, vol. 1. IEEE, 2003, pp. I–I.

[23] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[24] C. Cieri, D. Miller, and K. Walker, "Fisher english training speech parts 1 and 2," *Philadelphia: Linguistic Data Consortium*, 2004.

[25] V. Teller, "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition," *Computational Linguistics*, vol. 26, no. 4, pp. 638–641, 2000.

[26] B. Castellano, "PySceneDetect," https://github.com/Breakthrough/PySceneDetect, 2020.

[27] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *Proc. ICCV*, 2017, pp. 192–201.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[29] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of classification*, vol. 1, no. 1, pp. 7–24, 1984.

[30] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *Proc. ICASSP*. IEEE, 2019, pp. 3965–3969.

[31] J. S. Chung and A. Zisserman, "Lip reading in profile," in *Proc. BMVC.*, 2017.

[32] A. B. Johnston and D. C. Burnett, *WebRTC: APIs and RTCWEB protocols of the HTML5 real-time web*. Digital Codex LLC, 2012.

[33] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[34] J. S. Chung, B.-J. Lee, and I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," in *Interspeech*, 2019, pp. 371–375.

[35] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. ACMM*, 2019.

[36] A. Dutta, A. Gupta, and A. Zisserman, "VGG image annotator (VIA)," http://www.robots.ox.ac.uk/ vgg/software/via/, 2016.

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

[38] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. ECCV*. Springer, 2006, pp. 531–542.

[39] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "Plda for speaker verification with utterances of arbitrary duration," in *Proc. ICASSP*. IEEE, 2013, pp. 7649–7653.

[40] 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." in *Interspeech*, 2018, pp. 2793–2797.