SEEHEAR: SIGNER DIARISATION AND A NEW DATASET

Samuel Albanie[1⋆] *Gül Varol*[1,2*] *Liliane Momeni*[1⋆] *Triantafyllos Afouras*[1]
Andrew Brown[1] *Chuhan Zhang*[1] *Ernesto Coto*[1] *Necati Cihan Camgöz*[3] *Ben Saunders*[3]
Abhishek Dutta[1] *Neil Fox*[4] *Richard Bowden*[3] *Bencie Woll*[4] *Andrew Zisserman*[1]

[1] Visual Geometry Group, University of Oxford, UK
[2] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[3] CVSSP, University of Surrey, Guildford, UK
[4] Deafness, Cognition and Language Research Centre, University College London, UK
https://www.robots.ox.ac.uk/~vgg/data/seehear/

ABSTRACT

In this work, we propose a framework to collect a large-scale, diverse sign language dataset that can be used to train automatic sign language recognition models.

The first contribution of this work is SDTRACK, a generic method for signer tracking and diarisation in the wild. Our second contribution is SEEHEAR, a dataset of 90 hours of British Sign Language (BSL) content featuring a wide range of signers, and including interviews, monologues and debates. Using SDTRACK, the SEEHEAR dataset is annotated with 35K active signing tracks, with corresponding signer identities and subtitles, and 40K automatically localised sign labels. As a third contribution, we provide benchmarks for signer diarisation and sign recognition on SEEHEAR.

Index Terms— Signer Diarisation, Sign Language Datasets

1. INTRODUCTION

Sign languages represent the natural means of communication of deaf communities. They are visual languages with grammatical structures that differ considerably from those of spoken languages; there is no universal sign language and unrelated sign languages are mutually unintelligible [37]. While there has been considerable progress in machine comprehension of spoken languages in recent years [18], automatic recognition of sign languages remains challenging [23]. A key obstacle is the scarcity of appropriate training data: sign language datasets are typically (i) orders of magnitude smaller than their spoken counterparts [3], and (ii) collected in constrained settings (e.g. lab conditions, TV interpreters), that do not reflect the complexity and diversity of "real-life" sign language interactions. For example, multiple signers often engage in conversation [9], which introduces the additional challenge of identifying *who* is signing *when*.

In this work, our goal is to address this issue by collecting and automatically annotating a large-scale, diverse dataset of signing under a broad range of conditions to allow training of strong sign language recognition models. This requires identifying (i) temporal segments *when* signing occurs, and (ii) *who* is signing as there may be several people present. Our first contribution is therefore to propose SDTRACK, a generic framework for signer detection, tracking and diarisation in the wild (Sec. 3). Our second contribution is to employ SDTRACK, together with a recently proposed automatic annotation technique [1] to collect the SEEHEAR dataset from TV broadcasts of the programme *SeeHear*, which is presented entirely in BSL. The show includes signing in diverse conditions (both indoors and outdoors in a variety of scenes) and contains a rich variety of interviews, monologues and debates. The signing content is accompanied by written English translations, obtained from broadcast subtitles. It was the first magazine programme series for the British deaf community, launched in 1981, and continues to be broadcast. The SEEHEAR dataset is annotated with active signing tracks, along with corresponding signer identities and subtitles, in addition to localised sign labels (Sec. 4). Finally, we provide benchmarks for the tasks of signer diarisation and sign recognition on SEEHEAR (Sec. 5).

Fig. 1. **The SEEHEAR dataset.** In this work, we propose a signer tracking and diarisation framework and automatically annotate 90 hours of *in-the-wild* sign language content.

2. RELATED WORK

Signer diarisation. Signer diarisation has received limited attention in the computational literature. Existing works have

*Equal contribution.

Fig. 2. The SDTRACK framework for diarisation comprises three modules: (left) *Multi-person tracker* forms person tracks and estimates the pose of each person, (middle) *Active signer detector* determines which (if any) of the tracked persons are actively signing, (right) *Signer re-identification* associates identities to each tracked person. See Sec. 3 for more details.

considered small-scale experiments (i.e. on four short videos in constrained conditions) using low-level motion cues [16, 17]. Several methods have been proposed for active signer detection (which aims to estimate whether a signer is active in a given video frame): [30] proposes an efficient online system appropriate for videoconferencing software, while [2] proposes a multi-layer RNN model which performs frame-level active signer detection. To the best of our knowledge, no prior work has considered signer diarisation at large-scale in unconstrained environments.

Sign language datasets. A number of sign language datasets have been proposed in the literature (see [23] for a survey of the predominant benchmarks). Key shortcomings of existing datasets include: a lack of diversity (in terms of signing environment, number of signer identities, or both), restricted domain of discourse [5, 24] (for example, weather broadcasts) and limited scale [3]. MSASL [22], WLASL [25], BSL-DICT [29] and BSL SignBank [14] cover a wide vocabulary, but are restricted to isolated signs. BSLCORPUS [35] provides fine-grained linguistic annotation of conversations and narratives, but is limited to pairs of signers under lab conditions. BSL-1K [1] provides a large-scale dataset of continuous signs, but is derived entirely from crops of inset public broadcast interpreters (occupying a fixed region of the screen) and lacks scene diversity and interaction.

3. DIARISATION

Task formulation. Given an unconstrained video containing an arbitrary number of signers, our objective is to answer the question: "*who* is signing *when*?" by parsing the video into spatio-temporal tracks of active signers together with their identities. Our task therefore represents the combination of two techniques: (i) *active signer detection* which seeks to answer the question "which person is signing?", and (ii) *signer identification* which addresses the question "who is this person?"

Concretely, given a sequence of video frames $\mathbf{x}_{1:T} = \{x_1, \ldots, x_T\}$ containing signing content, the objective is to identify active signers at each frame, x_t, and to provide bounding boxes representing projections of each individuals' 3D signing space on the image plane.

Proposed Framework: SDTRACK. Our approach, SDTRACK (an abbreviation of Signer Diarisation and Tracking) comprises three modules: (1) a multi-person tracker, (2) an active signer detector, and (3) a signer re-identification model (see Fig. 2 for an overview). The role of each of these modules are described next.

Multi-Person Tracking. The objective of the tracker is to localise each person appearing in the footage and track their trajectories across frames. The tracker output consists of a collection of spatio-temporal bounding box trajectories $\mathcal{T} = \{\tau_i : \tau_i \in \mathbb{R}^{l_i \times 4}, i \in \{1, \ldots, N\}\}$, where N represents the total number of trajectories, l_i denotes the length of trajectory i, and τ_i specifies an array of bounding box locations in pixel coordinates.

Active Signer Detector. For each trajectory τ_i delivered by the tracker, the role of the active signer detector is to produce a vector, $y_i = \{0, 1\}^{l_i}$, which associates to each trajectory bounding box a label of 1 to indicate that the person contained within it is actively signing, and 0 otherwise.

Tracked person Re-identification (Re-ID) module. The role of this module is to group the trajectories $\mathcal{T}$ delivered by the tracker by their identity. To each track, τ_i, the model assigns an identity label $\eta_i \in [0, \ldots, M_{\mathcal{T}}]$, where $M_{\mathcal{T}}$ denotes the estimated total number of identities in $\mathcal{T}$.

Finally, the modules are combined by segmenting each person track τ_i into the sequences of contiguous active signing frames indicated by y_i and assigning to each segment its corresponding track identity identity label η_i.

Implementation. We instantiate the multi-person tracking module with the robust pose-based tracker recently proposed by [32]. In brief, this method employs a YOLOv3 [33] object detector, trained for person detection on the MS COCO dataset [26] to produce initial bounding box proposals at a fixed frame interval. Human keypoint estimation is then performed within the proposals over consecutive frames by recursively updating the bounding box such that it encloses the keypoints. These keypoints are fed to a two-layer Siamese graph convolutional network (which is trained to estimate the similarity of two poses via a learned metric in 128-dimensional feature space) that is used together with spatial consistency constraints to link box-enclosed poses into tracks.

While SDTRACK is agnostic to the implementation of the tracker, the pose-based tracking described above has the auxiliary benefit of estimating a set of human keypoints for each trajectory bounding box which can be efficiently re-used by the active signer detector module. We therefore implement the *active signer detector* as a binary classifier which takes as input the keypoint trajectories for a track and predicts its labels y_i. We explore two variants of the active signer detector: (1) a simple frame-level decision tree heuristic which predicts an active signer whenever the estimated forearm-to-upper-arm angle is less than a fixed threshold (1.1 radians—determined through visual inspection on example data); (2)

An eight-layer convolutional network which consumes temporal sequences of estimated keypoint confidences and locations (normalised with respect to their enclosing bounding box) and produces frame-level active signer labels. The latter model, implemented with the framework provided by [13], is "bootstrapped" by training on the predictions of the former across the training set of SEEHEAR (described in Sec. 4).

The *Signer Re-ID module* employs face verification for inter-track Re-ID (linking identities across tracks) and relies on the person tracker for *intra-track* Re-ID (identity is assumed consistent within a person track). Inter-track Re-ID comprises a three stage pipeline. (1) Faces are detected using the RetinaFace architecture [11], with Mobilenet0.25 [20] as the backbone network, trained on the WIDER FACE dataset [41]. (2) Face-identity discriminating embeddings are extracted from the face detections using an SENet-50 [21] architecture (pre-trained on the MS-Celeb-1M [19], and fine-tuned on VGGFace2 [6]). (3) As with the person tracker, face detections of the same identity in consecutive frames are linked together into face-tracks, using a combination of IoU (intersection over union) and similarity between face embeddings to link detections. For each face-track, the embeddings from the constituent detections are average-pooled across the track and L2-normalised, giving a single face-track embedding.

We evaluate these design choices and ablate their effect on diarisation performance in Sec. 5.

4. THE SEEHEAR DATASET

Dataset Description. The SEEHEAR dataset comprises 90 hours of British Sign Language (BSL) content produced by more than 1000 signers. It is sourced from publicly available broadcast archive footage of the long-running UK BBC show, *SeeHear*. We provide 35K active signer tracks and 40K automatically localised instances of specific signs for this data.

The partitions of the dataset are shown in Tab. 1. We note that in a number of episodes for the first several seasons of the show, presenters used "Sign-Supported English" (SSE)[1], which differs from true BSL in terms of grammar and form of sign production [39, 40]. These data are excluded in this work from algorithmic performance evaluation but may nevertheless be of interest to the sign recognition community. We therefore estimate the presence of SSE in each signing track automatically with the method of [8], and provide this as additional metadata.

The *Train*, *Val*, *Test (public)* and *Test (private)* partitions (summarised in Tab. 1) are formed by assigning the earliest 80 episodes (for which SSE is more prevalent) to *Train*, then randomly assigning the remaining episodes across splits in the given proportions. The latter private test partition will be withheld to support the future use of the dataset as a benchmark for sign language tasks.

[1]We note that SSE lacks a universally agreed upon definition [4]. In our work, it refers to simultaneous production of both sign and speech.

	Num. vids	Track duration (in hours)	Sparse annos	Num. tracks (of which SSE)
Train	275	75.7	36.7K	28.7K (10.6K)
Val	17	3.7	1.4K	1.6K (0.5K)
Test (public)	27	6.6	2.4K	2.7K (0.8K)
Test (private)	25	5.7	2.1K	2.5K (0.8K)
Total	344	91.7	5,352 (6,945)	35.6K (12.8K)

Table 1. Statistics for the partitions of the SEEHEAR dataset. See Sec. 4 for further details.

Dataset Pipeline. To construct the SEEHEAR dataset, SDTRACK was employed as part of a multi-stage dataset pipeline, described next.

1. Source footage was obtained from 344 episodes of the TV show, *SeeHear*, from public UK library sources [38], enabling its use for non-commercial research.

2. OCR subtitle extraction. Subtitle text is first detected using the PixelLink algorithm [10, 27], and then recognised via the matching method of [42] trained on synthetic text lines. For each frame sequence containing a subtitle text, the text recognised on each frame is then analysed with an English transformer language model [31] trained with the Fairseq sequence modelling toolkit. The text with the lowest perplexity is chosen as the subtitle for the whole frame sequence.

3. Signer tracking and diarisation. Using the SDTRACK framework, tracks were obtained from all episodes and diarised into active signer segments. Each active signer is associated with the corresponding subtitles produced by the previous stage. *4. Annotations from mouthings.* Within the active signer segments produced by SDTRACK, we apply the sign spotting method proposed by [1] using the improved visual-only keyword spotting model of Stafylakis et al. [36] from [28] (referred to in their paper as "P2G [36] baseline"). To generate the list of candidate signs for spotting, the subtitles are first text-normalised using the method of [15] (this converts non-standard words such as numbers and date abbreviations to a canonical written form, e.g. "35" is mapped to "thirty five"). Signing tracks are then queried with each word present in the normalised subtitles that possesses four or more phonemes. Following [1], sparse annotations are retained for all spotting predictions which are assigned a posterior probability of 0.5 or greater by the visual keyword spotting model. These localised sign annotations allow SEEHEAR to be used as a sign recognition benchmark.

The outputs of this fully automatic pipeline, which comprise SEEHEAR with active signing tracks along with subtitle sentences and sparse sign annotations, are summarised in Tab. 1. In addition to the above, the dataset also includes a set of 4,478 sparse annotations (distributed across all dataset partitions) that have been manually verified as correct by annotators (described in more detail in Sec. 5).

	DER↓	JER↓	NMI↑
SDTRACK (w/o ASD,Re-ID)	71.1	53.8	0.76
SDTRACK (w/o Re-ID)	62.9	51.3	0.79
SDTRACK (w/o ASD)	25.3	41.9	0.86
SDTRACK	**12.2**	**29.8**	**0.93**
SDTRACK with ASD heuristic	12.3	29.9	0.93
SDTRACKwith bootstrapped ASD ConvNet	12.2	29.8	0.93

Table 2. Signer diarisation performance on annotated portion of SEEHEAR public test set. (Upper) the effect of removing the *Active Signer Detection (ASD)* and *Re-ID* modules from the SDTRACK framework. (Lower) The influence of different ASD methods on overall diarisation performance.

Pretraining	Finetuning	per-instance top-1	per-instance top-5	per-class top-1	per-class top-5
BSL-1K [1]	-	21.2	38.4	22.3	38.6
Kinetics	SEEHEAR$_{m.5}$	52.4	68.7	39.0	54.2
BSL-1K	SEEHEAR$_{m.7}$	63.5	80.1	51.9	73.2
BSL-1K	SEEHEAR$_{m.5}$	**67.4**	**82.1**	**58.0**	**75.4**

Table 3. Sign recognition on the SEEHEAR benchmark: We provide benchmark results for sign classification on the manually verified public test set of SEEHEAR.

5. EXPERIMENTS

In this section, we report on using ground truth annotations to evaluate components of the SDTRACK framework, validate the automatic sparse annotations, and assess automatic sign recognition performance on the sparse annotations.

Signer Diarisation: benchmark and metrics. To evaluate diarisation performance, six videos from the public test partition were manually annotated with active speaker and identity labels (resulting in a total of 104 turns of signing across 43 minutes of video). We report three metrics to compare performance: Diarisation Error Rate (DER) (which is formed from the sum of missed signing, false alarm signing, and signer misclassification error rates), Jaccard Error Rate (a metric recently introduced for the DIHARD II diarisation challenge [34] which represents the ratio of intersections to unions between predicted and ground truth signer turns) and Normalised Mutual Information (NMI) between predicted and ground truth signer turns.

Ablations. In Tab. 2 (upper), we first perform ablations by removing various components of the SDTRACK pipeline. The results confirm that each module plays an important role in diarisation performance. In Tab. 2 (lower), we fix the *Multi-Person Tracking* and *Signer Re-identification* modules of the SDTRACK framework and evaluate the two proposed variants for the active signer detection module. We observe that the ConvNet, trained on the predictions of the crude decision tree heuristic, performs similarly to the baseline. However, given its marginal improvement, we use this model for the final SDTRACK implementation.

Sparse annotation validation. We next provide an assessment of the quality of the sparse annotations listed in Tab. 1 that were obtained using mouthing cues through visual keyword sign spotting (using the method of [1]). To quantify annotation quality, we extended the open-source VIA tool [12], to allow human experts in sign language to assess the correctness of approximately 10K sign spotting predictions. Of these, 60.4% of the predictions were marked correct, validating the effectiveness of the automatic annotation approach. We then further investigated a random sub-sample of 100 annotation errors to assess their causes. We found that 74% of annotation errors could be attributed to mistakes made by the visual keyword spotting model, and 26% resulted from failures in the tracking and diarisation pipeline to correctly identify the currently active signer in the scene (to whom the content of the subtitles corresponds), highlighting the room for further improvement of the latter.

Automatic sign recognition. As described in Sec. 4, besides diarisation, SEEHEAR can be used for sign language recognition. Here, we provide benchmark results for classifying signs given short temporal intervals around sparse annotations within active signer tracks. Specifically, the input is a 20-frame video cropped around the active signer. We determine a vocabulary of 4321 signs from the train split, removing words for which all mouthing annotation confidences are below 0.8. For the training set (combining Train+*SSE* splits), we use 36K automatic mouthing annotations. For testing, we use the 307 manually verified signs which fall within the public test partition spanning a vocabulary of 124 signs. Tab. 3 reports the recognition performance for several I3D [7] models with or without fine-tuning on the SEEHEAR training set. First, we apply the state-of-the-art model of [1], which is trained for 1064 sign categories on BSL-1K, covering our test vocabulary. We obtain 21.2% accuracy, which highlights the difficulty of generalisation to in-the-wild SEEHEAR data. Second, we finetune this model for the more challenging task of 4321-way classification and observe significant gains (67.4%). Third, similar to [1], training with a less noisy but smaller set of annotations by filtering the mouthing confidence at 0.7 (SEEHEAR$_{m.7}$) degrades the performance (63.5%). Finally, sign recognition pretraining on BSL-1K gives a significant boost over action recognition pretraining with Kinetics (52.4% vs 67.4%).

6. CONCLUSION

In this work, we introduced the SDTRACK framework for signer tracking and diarisation in the wild. We used SDTRACK to collect and annotate a large-scale collection of BSL signing content in diverse conditions to produce the SEEHEAR dataset. We also provide several diarisation and recognition baselines on the introduced dataset to underpin future research on sign language understanding in the wild.

Acknowledgements. This work was supported by EPSRC grant ExTol. The authors would like to thank Himel Chowdhury for his assistance with sign annotation.

References

[1] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *ECCV*, 2020.

[2] M. Borg and K. P. Camilleri, "Sign language detection "in the wild" with recurrent neural networks," in *ICASSP*, 2019.

[3] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. K. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Morris, "Sign language recognition, generation, and translation: An interdisciplinary perspective," *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019.

[4] British Deaf Association, "The Difference between BSL & SSE," 2017. [Online]. Available: https://bda.org.uk/the-difference-between-bsl-sse/#_ftn5

[5] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural Sign Language Translation," in *CVPR*, 2018.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *FG*, 2018.

[7] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *CVPR*, 2017.

[8] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[9] J. Coates and R. Sutton-Spence, "Turn-taking patterns in deaf conversation," *Journal of Sociolinguistics*, 2001.

[10] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *AAAI*, 2018.

[11] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *ArXiv*, vol. abs/1905.00641, 2019.

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

[13] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3575–3584.

[14] J. Fenlon, K. Cormier, R. Rentelis, A. Schembri, K. Rowley, R. Adam, and B. Woll, "Bsl signbank: A lexical database and dictionary of british sign language (first edition)," 2014.

[15] E. Flint, E. Ford, O. Thomas, A. Caines, and P. Buttery, "A text normalisation system for non-standard English words," in *W-NUT*. ACL, Sep. 2017, pp. 107–115.

[16] B. G. Gebre, P. Wittenburg, T. Heskes, and S. Drude, "Motion history images for online speaker/signer diarization," in *ICASSP*, 2014.

[17] B. Gebrekidan Gebre, P. Wittenburg, and T. Heskes, "Automatic signer diarization-the mover is the signer approach," in *CVPRW*, 2013.

[18] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[20] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *ArXiv*, vol. abs/1704.04861, 2017.

[21] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks." *IEEE TPAMI*, 2019.

[22] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv preprint arXiv:1812.01053*, 2018.

[23] O. Koller, "Quantitative survey of the state of the art in sign language recognition," *arXiv:2008.09918*, 2020.

[24] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *CVIU*, 2015.

[25] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *WACV*, 2020.

[26] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[27] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *ECCV*, 2018.

[28] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman, "Seeing wake words: Audio-visual keyword spotting," *BMVC*, 2020.

[29] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, "Watch, read and lookup: learning to spot signs from multiple supervisors," in *ACCV*, 2020.

[30] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-Time Sign Language Detection using Human Pose Estimation," in *ECCVW (SLRTP)*, 2020.

[31] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook fair's wmt19 news translation task submission," *ArXiv*, vol. abs/1907.06616, 2019.

[32] G. Ning, J. Pei, and H. Huang, "Lighttrack: A generic framework for online top-down human pose tracking," in *CVPRW*, 2020.

[33] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767*, 2018.

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

[35] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, and K. Cormier, "Building the British sign language corpus," *Language Documentation & Conservation*, vol. 7, 2013.

[36] T. Stafylakis and G. Tzimiropoulos, "Zero-shot keyword spotting for visual speech recognition in-the-wild," in *ECCV*, 2018.

[37] R. Sutton-Spence and B. Woll, *The Linguistics of British Sign Language: An Introduction*. Cambridge University Press, 1999.

[38] University of Bristol Library, "Library archives." [Online]. Available: https://bris.on.worldcat.org/external-search?queryString=see%20hear

[39] R. Whitehead, N. Schiavetti, B. Whitehead, and D. Metz, "Temporal characteristics of speech in simultaneous communication." *Journal of speech and hearing research*, vol. 38 5, pp. 1014–24, 1995.

[40] R. B. Wilbur and L. Petersen, "Modality interactions of speech and signing in simultaneous communication," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. 200–212, 1998.

[41] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016.

[42] C. Zhang, A. Gupta, and A. Zisserman, "Adaptive text recognition through visual matching," in *ECCV*, 2020.