GestSync: Determining who is speaking without a talking head
Supplementary Material

Sindhu B Hegde
sindhu@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

We report additional experiments and ablation studies. Please also look at the demo video, where we show results on multiple samples, including in-the-wild examples.

1 Ablation Studies

We perform ablation experiments using our best keypoint-vector based model. The results are for the LRS3 *validation* set [1], and are used to choose the best model parameters. The results in the main paper are for the LRS3 *test* set.

1.1 Temporal window

We provide justification for using a temporal window of $T = 25$ frames (1-second of audio and video segments) as input to our model in Table 1. Note that T specifies the input to the model, whereas $F = [25, 50, 75, 100]$ specifies the number of frames used when averaging the audio-visual similarity score. Our model consistently improves in performance as the length of the input segment increases, which aligns with our intuition that longer context windows are beneficial due to the sparse correlation between gestures and speech. From Table 1, it is evident that using only 5 frames leads to a significant deterioration in performance. When using a longer window of $T = 50$ frames, the performance is similar to that of $T = 25$ frames, with only a minor improvement. This behavior can be attributed to the limited availability of training data for longer input segments. Since we utilize a contrastive loss framework where shifted versions of the same video act as negative samples, obtaining sufficiently long videos for sampling negatives becomes challenging. Moreover, as explained in Section 3.4 of the main paper, our negative samples need to be at least 1 second away from the positive sample, further limiting the sampling process for longer videos when training the model with $T \geq 50$ frames.

Table 1: Synchronization performance variation on LRS3 val. The first column specifies the input window (the number of frames input to the GestureSync network). The variation across the columns specifies the number of frames used to average the score. Video and audio are sampled at 25 Hz. Longer input windows enable the model to capture the temporal context and effectively learn the gesture-speech synchronisation.

Temporal Window	**25**	**50**	**75**	**100**
5	32.1	39.5	46.2	51.7
15	40.4	48.9	57.6	63.2
25	43.2	51.5	58.8	64.1
50	44.5	51.7	58.9	64.3

1.2 Using additional hand keypoints

We show the effect of using additional hand keypoints in Table 2. Specifically, we utilize the hand keypoints (extracted from Mediapipe [3]) along with the pose keypoints. This gives a total of 64 keypoints which act as input to the GestureSync model (22 pose keypoints + 21 keypoints for each hand). Using more keypoints gives us an improvement across all the averaging windows F, except for the largest window of 100 frames.

Table 2: Comparison of using additional hand keypoints along with pose keypoints. Synchronization performance variation on LRS3 val.

Method	**25**	**50**	**75**	**100**
pose (22 kps)	43.2	51.5	58.8	64.1
pose + hands (64 kps)	45.3	52.9	59.0	62.2

2 Additional Experiments

2.1 Evaluation on LRS3-Lang dataset

In addition to the evaluation on the LRS3 dataset [1] shown in the main paper (Section 4.3), we also assess the model on the LRS3-lang dataset [2]. LRS3-lang is a multi-lingual dataset comprising 12 different languages with a total of over 1300 hours of video data. We obtained this dataset for our evaluation from the authors (since the data has not yet been publicly released). Note that the pre-processing used in the provided body-crops data (including resolution of videos, and the bounding-box used to create the body crops) is entirely different from that of LRS3. Thus, we fine-tune our models (for 5 epochs) to adapt to the different pre-processing settings. Since the official train-test splits are not yet provided for this dataset, we randomly sample $\sim 3\%$ videos without any speaker overlaps to create our test set (we will release the splits). The distribution of train and test sets across multiple languages is shown in Figure 2. The evaluation on a more challenging, multi-lingual LRS3-lang dataset highlights the capabilities of the GestureSync model and shows that it is not limited by language barriers.

Table 3 shows the results of the models using different input representations on the LRS3-lang dataset [2]. In-line with the LRS3 results shown in the main paper (Table 2), our RGB-based Transformer model achieves the best performance. It is worth noting that the synchronisation accuracy on LRS3-lang is very similar to that of LRS3, despite the fact that

LRS3-lang is a much harder dataset with a wider variations in terms of speakers, languages, and resolution. This indicates the abilities of our model to work effectively in challenging settings. inputs.

Table 3: Performance comparison of gesture synchronisation accuracy (%) averaged over a given number of frames F on the LRS3-lang dataset [2].

Method	25	50	75	100
Ours - RGB	44.8	56.0	67.9	76.4
Ours - Keypoint-image	37.2	47.4	52.3	58.9
Ours - Keypoint-vector	40.1	46.7	54.5	62.5

2.2 Do gestures vary with the language spoken?

We investigate the variations in gesture-speech correlations across different languages (and hence to some extent nationalities) worldwide. To conduct this analysis, we utilize the LRS3-lang dataset, and leverage the provided language labels to categorize the speakers accordingly. We specifically evaluate on eight languages, excluding Polish, Turkish, Arabian and Greek due to the limited availability of test data in these languages (see Figure 2 (b)).

In Table 3, we compute the synchronisation accuracy (averaged over F frames, where $F = [25, 50, 75, 100]$) for each language category individually. It is evident that certain languages exhibit a stronger and more explicit correlation, such as Italian, Portuguese, Spanish, and French. Conversely, the correlation between gestures and speech is less pronounced for Korean and Japanese speakers, showcasing an opposite trend.

Table 4: We study how the synchronization performance of the gesture-speech model varies across different languages using the LRS3-lang dataset. While some languages like Italian, Portuguese, Spanish have stronger gesture-speech correlations, languages like Japanese and Korean have weaker links.

Language	25	50	75	100
german	40.9	45.5	68.2	75.0
portuguese	44.9	60.5	71.5	73.8
spanish	46.8	59.5	72.2	80.3
french	47.1	58.9	65.5	78.2
russian	42.2	54.2	63.1	72.0
japanese	39.7	42.8	55.6	66.7
italian	**49.3**	**61.6**	**74.0**	**82.3**
korean	35.1	38.9	49.6	64.9

2.3 Sensitivity to speaker attributes

We analyse the behaviour of the GestureSync model on the different speaker attributes such as gender and age of the speakers. For gender classification and age estimation, we use OpenCV-based public implementation[1] and obtain the labels on the LRS3 test set [1].

Table 5 demonstrates that the synchronization performance is higher in the female category compared to the male category. This observation suggests that there may be inherent

[1] https://github.com/smahesh29/Gender-and-Age-Detection

differences in the gestures and speech patterns of male and female speakers. Further investigation and analysis of these gender-related differences could provide valuable insights into the dynamics of correlation between gestures and speech. Table 5 also reveals that there is no significant difference in performance across different age groups. This observation could be attributed to the nature of the LRS3 dataset, which predominantly consists of videos from trained TED speakers. The dataset's composition of experienced speakers might mitigate any potential impact of age on the performance of speech-gesture synchronization. Therefore, the age of the speakers does not appear to be a significant contributing factor in determining the level of synchronization achieved in this context.

Table 5: Effect of the speaker attributes such as gender and age on model's synchronization performance on LRS3 test set [1]. The GestureSync network performs better for female category, whereas the performance remains consistent across different age-groups.

Attribute	**Class**	**25**	**50**	**75**	**100**
Gender	Female	42.9	50.5	59.2	64.4
	Male	41.1	48.2	56.2	61.2
Age	< 25	42.8	48.0	56.3	62.9
	$25 - 60$	42.1	49.4	56.1	62.4
	> 60	42.9	49.1	57.1	63.2

2.4 Improving the keypoint-vector representation

As explained in Section 4.3.1 of the paper, one of our long-term goals is to bridge the gap between RGB and keypoint-vector representations. We perform several further experiments as explained below to investigate to what extent we can boost the keypoint-vector model's performance.

2.4.1 Data Augmentation

One of the techniques which has proven to be beneficial in various image and video processing tasks is data augmentation. Strategies such as translation, flipping, rotation, and scaling are used to improve the robustness of the model, thus enhancing the performance during inference. Following these traditional techniques, we too apply data augmentation to our keypoint-vector representation network. Specifically, we apply the following augmentations: (i) Shifting – Randomly shift the x and y co-ordinates of the keypoints in the range of $[-50, 50]$, (ii) Rotation – Rotate all the keypoints by an angle in the range of $[-10, 10]$, (iii) Scaling – Scale the keypoints randomly in the range of $[0.7, 1.3]$. Table 6 shows the results of augmentations. We can observe that using data augmentation techniques results in further boosts in the performance.

Table 6: Comparison of adding data augmentation to keypoint-vector representation model on LRS3 test set. Adding augmentation helps in improving the performance.

Method	**25**	**50**	**75**	**100**
W/o augmentation	41.7	49.8	58.1	62.7
With augmentation	43.1	51.2	59.5	64.2

2.4.2 Using head motion information

When the person is speaking, head movements convey vital natural motion information alongside gestures. While we opt *not* to incorporate lip motion data in our study, we have the flexibility to make use of head motion when it is accessible. We extract the face keypoints using Mediapipe and consider the face-oval/head keypoints along with the pose keypoints for this experiment. Utilizing all face-oval keypoints could potentially introduce lip motion information from the lower jaw regions, so to completely avoid any lip-related input, we conduct another experiment focusing solely on the upper head (above the ears). The outcomes are presented in Table 7. Remarkably, we achieve an almost perfect score of 95.7% with a 100-frame average when utilizing all head keypoints. As previously mentioned, this performance can be attributed to potential lip motion leakage. Notably, the performance of pose combined with upper head keypoints (the last row in the table) demonstrates a substantial enhancement compared to using only pose keypoints, highlighting the significant role of head motion in determining synchronisation.

Table 7: Performance comparison of utilizing the head motion information in determining synchronisation (on LRS3 test set).

Method	**25**	**50**	**75**	**100**
pose (22 kps)	41.7	49.8	58.1	62.7
pose + head (58 kps)	77.6	88.6	94.3	95.7
pose + upper head (43 kps)	49.8	60.9	70.1	76.2

3 Visualisation

3.1 RGB input representation: Masked frames

Figure 1 demonstrates the input masked frames for our RGB representation based model. To avoid using face and lip motion information, we mask the face region as shown in the figure.

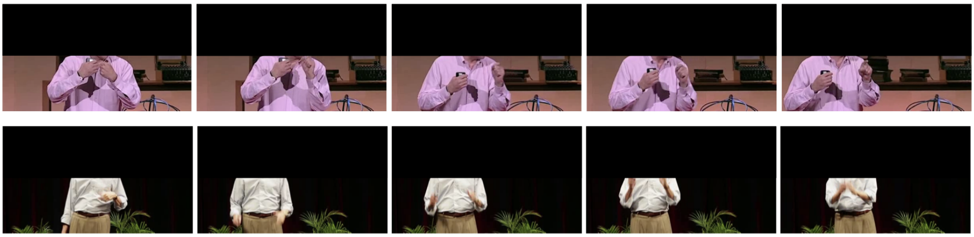

Figure 1: Visualisation of input frames for RGB representation based network for gesture synchronisation. The face is masked to avoid using any lip motion information.

3.2 Gesture variation

Figure 3 illustrates the variation of gestures across different speakers, highlighting that not all speakers exhibit expressive gestures that can be readily associated with their speech. These variations pose a challenge for our task. However, by providing a longer temporal context, the model can effectively aggregate subtle cues and make confident predictions regarding the synchronization between speech and gesture.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Now you're speaking my language: Visual language identification. In *INTERSPEECH*, 2020.

[3] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

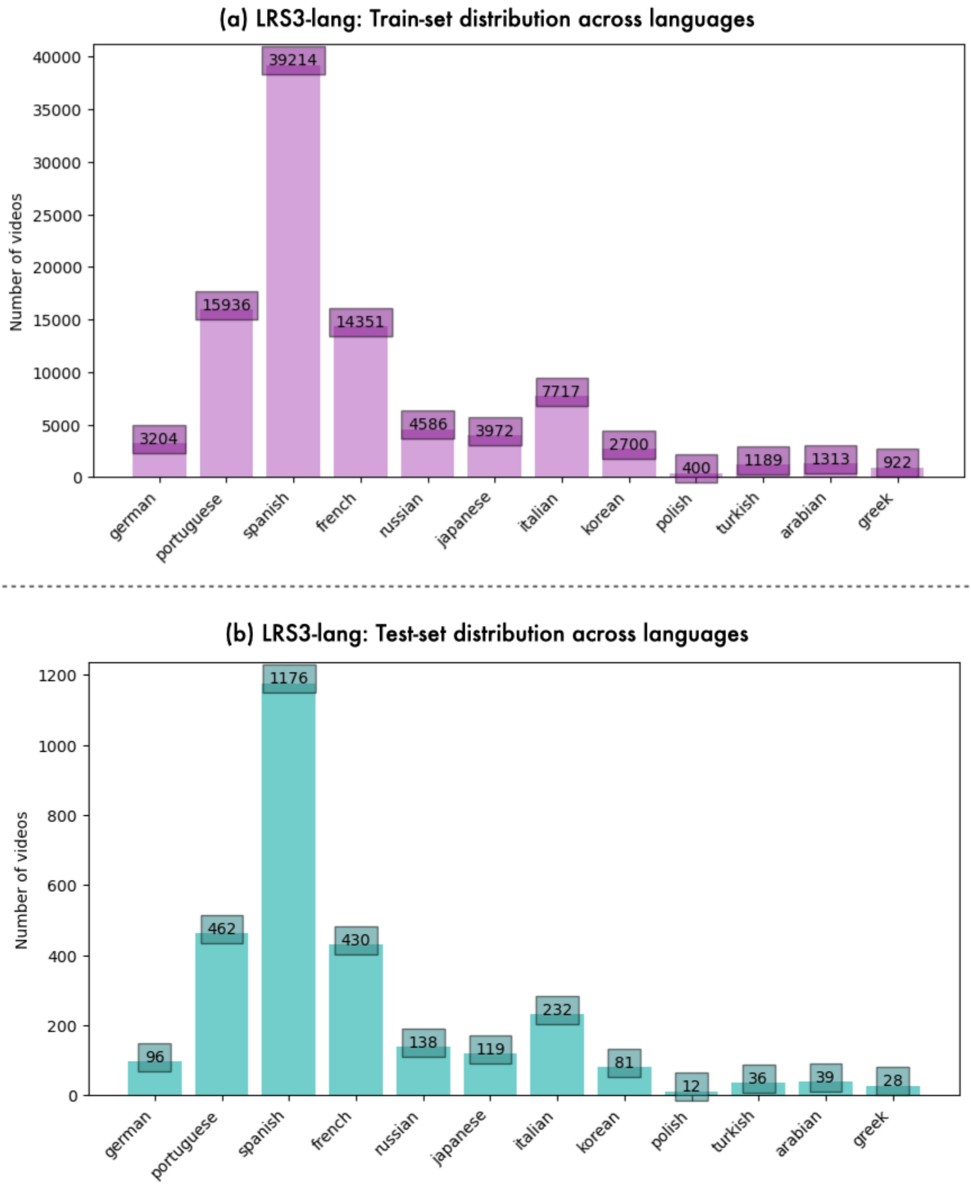

Figure 2: We show the distribution of languages in the training and test sets of LRS3-lang dataset [2].

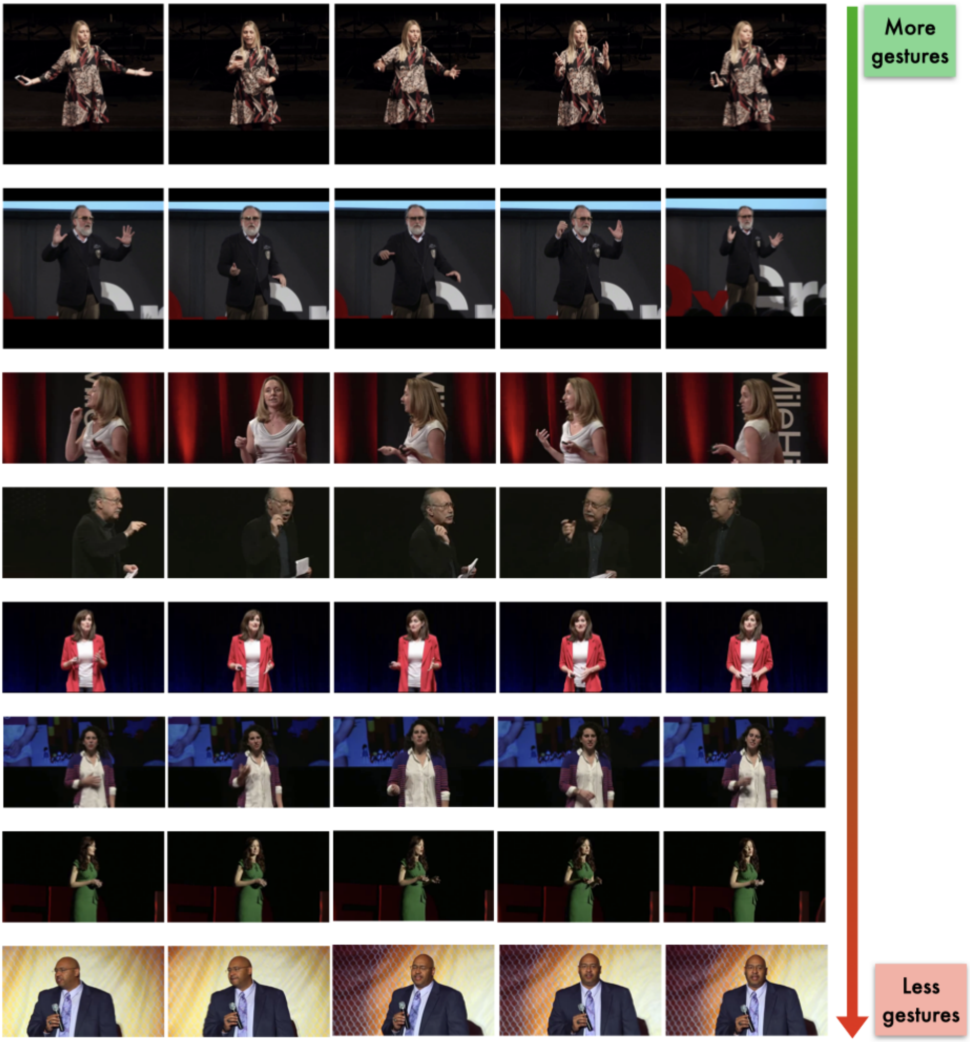

Figure 3: Variation of gestures across different speakers. A few speakers exhibit clear and expressive gestures (top rows) while others exhibit less prominent gestures (bottom rows). This diversity highlights the inherent challenges associated with our task.