Supplementary Material
Understanding Co-speech Gestures in-the-wild

Sindhu B Hegde[*], K R Prajwal[*], Taein Kwon, Andrew Zisserman
Visual Geometry Group, Dept. of Engineering Science, University of Oxford
{sindhu, prajwal, taein, az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/jegal

1. Most Gestured Words

In Figure 1, we show the most commonly spotted gestured words that are spotted by JEGAL on the AVS-Spot test set: pointing gestures (you, my, we), adjectives/adverbs (little, open, whole, gigantic, broad), direction words (forward, here, below) and numbers (one, two, first).

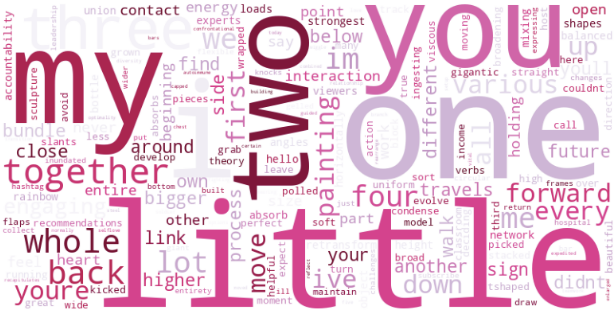

Figure 1. Word cloud for the most commonly gestured words.

2. Additional Evaluations and Analysis

2.1. Gesture Word Spotting: Evaluation in challenging conditions

Our evaluation set (constructed from AVSpeech) includes a diverse range of samples: (i) non-frontal videos, (ii) varying lighting conditions, (iii) a wide variety of speakers, and (iv) conversational videos (from which we extract segments featuring a single speaker). In this section, we specifically benchmark the performance of JEGAL on these challenging subsets. We label the AVS-Spot dataset with new metadata: (i) lighting conditions (dim, medium, bright), and (ii) speaker poses (frontal vs. non-frontal). Fig 2 illustrates the diversity of the test set.

Table 1 reports the spotting accuracy across these subsets. We find that the model performs best on brightly lit videos, with similar accuracy for dim and medium lighting.

Table 1. Evaluation in challenging conditions: JEGAL outperforms prior models in all settings.

Method	Lighting			Speaker pose	
	Dim	Medium	Bright	Frontal	Non-frontal
GestSync [4]	9.67	22.92	8.33	21.59	17.80
GestDiffuClip [2]	15.6	19.7	21.8	19.35	19.90
JEGAL (Ours)	61.29	62.58	77.77	62.76	68.49

2.2. Effect of Modality dropping

We present the impact of dropping text and audio modalities at varying rates on the spotting task in Table 2. A drop rate of 30% means that during training, either text or audio is randomly dropped in 30% of the batch samples. Dropping the modalities at 50% performs the best across all inference-time settings.

Table 2. Dropping modalities evenly during training works best.

Drop %	Accuracy ↑		
	T	A	TA
30%	52.2	38.6	63.2
50% (JEGAL)	61.0	41.8	63.6
70%	61.3	42.2	62.6

2.3. Computational efficiency

Table 3 shows the inference time (averaged across ten runs) for a 5-second input on a single NVIDIA $V100$ GPU. Our model can process ≈ 52 frames per second, indicating that the inference is quite fast but is not streaming-capable yet, as the bidirectional transformer attends to all future frames provided as context.

Table 3. Model parameters and inference time analysis for 5s input.

	Visual Enc.	Text Enc.	Audio Enc.	Total
inf. time (sec) ↓	1.84	0.18	0.07	2.09

2.4. Where does the model focus on?

We visualize the activation maps of the visual features of JEGAL to see which spatial region of the video the model focuses on. In Fig 3, we see that the model focuses on the hand gestures.

[*] equal contribution

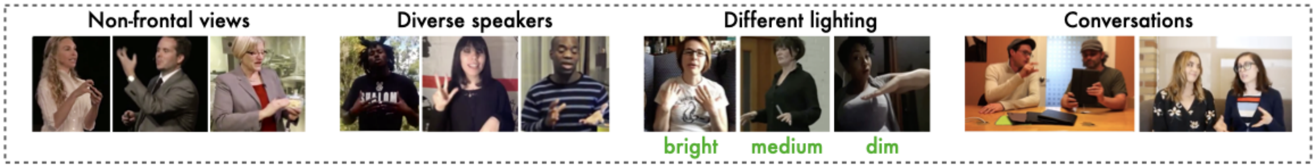

Figure 2. The AVS-Spot test set is quite diverse – some examples are shown above. Additionally, we annotate the clips in AVS-Spot for frontal/non-frontal views and lighting and analyze the performance on these individual subsets.

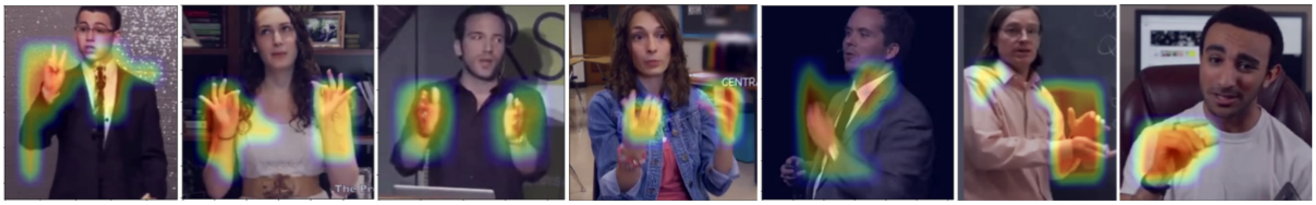

Figure 3. We plot the activation maps of the visual features of JEGAL. We can see that JEGAL focuses strongly on the hand gestures.

3. Model Details

In Table 4, we provide detailed description of the model architecture. The code and models to have been released to support future research.

4. Dataset Visualization

In Figure 4, we present examples from our manually annotated AVS-Spot test set (curated from the publicly available AVSpeech test dataset [3]), designed to evaluate downstream gesture spotting performance. As shown, the dataset includes a diverse collection of unique words, carefully curated to ensure clear and contextually appropriate gestures. For instance, in row-1, the word "little" is accompanied by a gesture where two fingers move close together to indicate a small size; in row-2, the speaker points backward to represent the word "back"; and in row-6, the fingers of both hands move in a distinctive pattern to indicate "hashtag".

5. Qualitative Results

In Figure 5, we show additional qualitative examples for gesture spotting. In the left text panel, the red-highlighted word represents the keyword to be spotted, as curated in the AVS-Spot test set. The word-labeled vertical columns, separated by yellow lines, indicate the word boundaries derived from speech-text alignment. JEGAL successfully spots most of these keywords, as shown by the red heatmaps. Notably, the boundaries may vary slightly since speakers often gesture and speak at slightly different times, highlighting the inherent challenges of our weakly-supervised gesture representation learning task.

In Figure 6, we present additional examples demonstrating that audio-based gesture spotting tends to focus on "stressed regions" in speech, unlike text-based spotting. This difference is evident from the audio and text heatmaps for each sample. In Figure 6, our model detects the stressed keywords "specific" and "respond", whereas the text-only model misses these words. Evidently, the audio-only model looks for word emphasis cues (indicated by high pitch) as such words are more likely to be gestured. This would be difficult to infer from text modality alone. These examples illustrate the advantages of leveraging audio cues for gesture spotting.

6. Limitations and Areas of Improvement

Our work is the first to tackle large-scale co-speech gesture understanding. We highlight some of the limitations of our approach here. One aspect the model struggles with is when there are limited gesture actions or hand movements that are unrelated to speech. Finally, given that we learn with only weak sequence-level supervision, the model can "find shortcuts" by focusing on simple rhythmic hand movements that occur in certain gestures classes like the beat gestures. This can affect the representation quality of iconic and deitic gestures that contain clear semantic meaning. While we still show that our models can spot such gestures, future works can focus on improving this imbalance in gesture classes.

7. Potential Negative Societal Impacts

While our research significantly contributes to advancing gesture understanding, there are some potential risks of surveillance, as the system could infer conversations from a distance by identifying words/phrases. Nonetheless, we believe the benefits outweigh these risks, as the technology enhances human-machine interaction by integrating non-verbal cues. According to the 55% rule [1], non-verbal communication constitutes 55% of overall communication. This highlights the importance of enabling machines to engage in holistic, natural interactions with humans by understanding non-verbal elements like gestures.

Table 4. Overview of the model architecture, detailing the input modalities, network components, and key parameters used in each stage of our framework.

Branch	Layer/Module	Input Shape	Output Shape
Visual Branch			
	Vision backbone	3 × T × 270 x 480	T × 1024
	Projection MLP		
	- Linear	T × 1024	T × 512
	- LayerNorm	T × 512	T × 512
	- ReLU	T × 512	T × 512
	- Linear	T × 512	T × 512
	Positional Encoding	T × 512	T × 512
	Transformer (N=6 layers)		
	- Self-Attention (h=8)	T × 512	T × 512
	- Feed Forward	T × 512	T × 512
	Output Projection	T × 512	T × 512
Text Branch			
	mRoberta Text backbone	W	W × 768
	Transformer (N=3 layers)		
	- Self-Attention (h=8)	W × 768	W × 768
	- Feed Forward	W × 768	W × 768
	Output Projection	W × 768	W × 256
Audio Branch			
	Melspectrogram Input	1 × 80 × 4T	-
	Conv2D + BN + ReLU		
	(k=5, s=1, p=2)	1 × 80 × 4T	32 × 80 × 4T
	Conv2D + BN + ReLU		
	(k=3, s=2, p=1)	32 × 40 × 2T	64 × 40 × 2T
	Conv2D + BN + ReLU		
	(k=3, s=2, p=1)	64 × 40 × 2T	128 × 20 × T
	Conv2D + BN + ReLU		
	(k=3, s=(3,1), p=1)	128 × 7 × T	256 × 7 × T
	Conv2D + BN + ReLU		
	(k=3, s=(3,1), p=1)	256 × 3 × T	256 × 3 × T
	Conv2D		
	(k=1, s=(3,1), p=0)	256 × 3 × T	256 × 1 × T
	Output Projection + reshape	256 × 1 × T	T × 256
Late Fusion			
	Encoded Features		
	- Visual	T × 512	-
	- Text + sub-word pooling	W × 256	W × 256
	- Audio + sub-word pooling	T × 256	W × 256

References

[1] 55% rule. `https://online.utpb.edu/about-us/articles/communication/how-much-of-communication-is-nonverbal/`. Accessed: 2024-11-21. 2

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023. 1

[3] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37, 2018. 2

[4] Sindhu B Hegde and Andrew Zisserman. Gestsync: Determining who is speaking without a talking head. In *Proc. BMVC*, 2023. 1

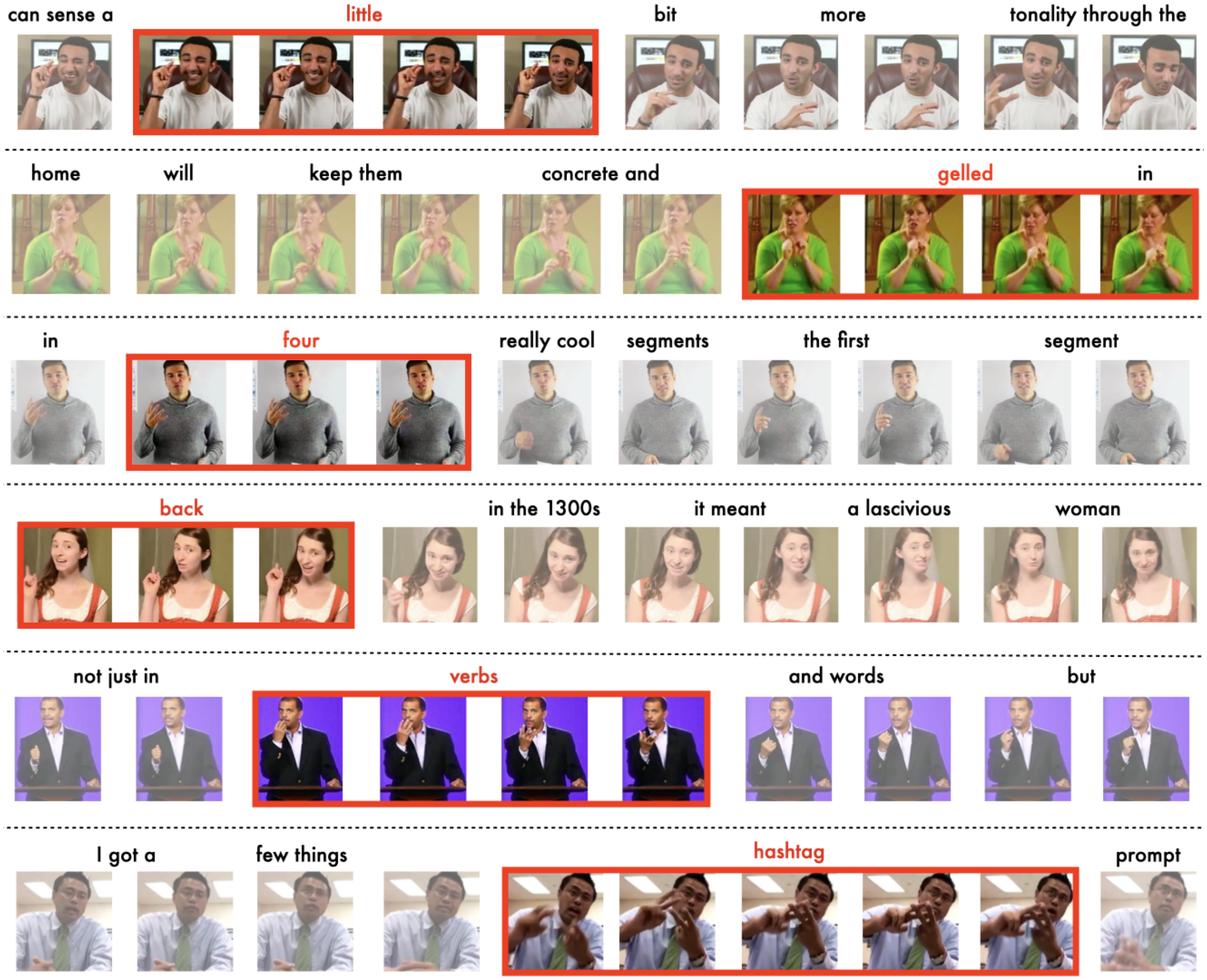

Figure 4. Visualization of the **AVS-Spot** dataset, showcasing video frames from different samples. Each row corresponds to a single video, with the highlighted keyword indicating the annotated gestured word for spotting. The figure illustrates the dataset's diversity, featuring a wide range of unique keywords, various speakers, and distinct gestures.

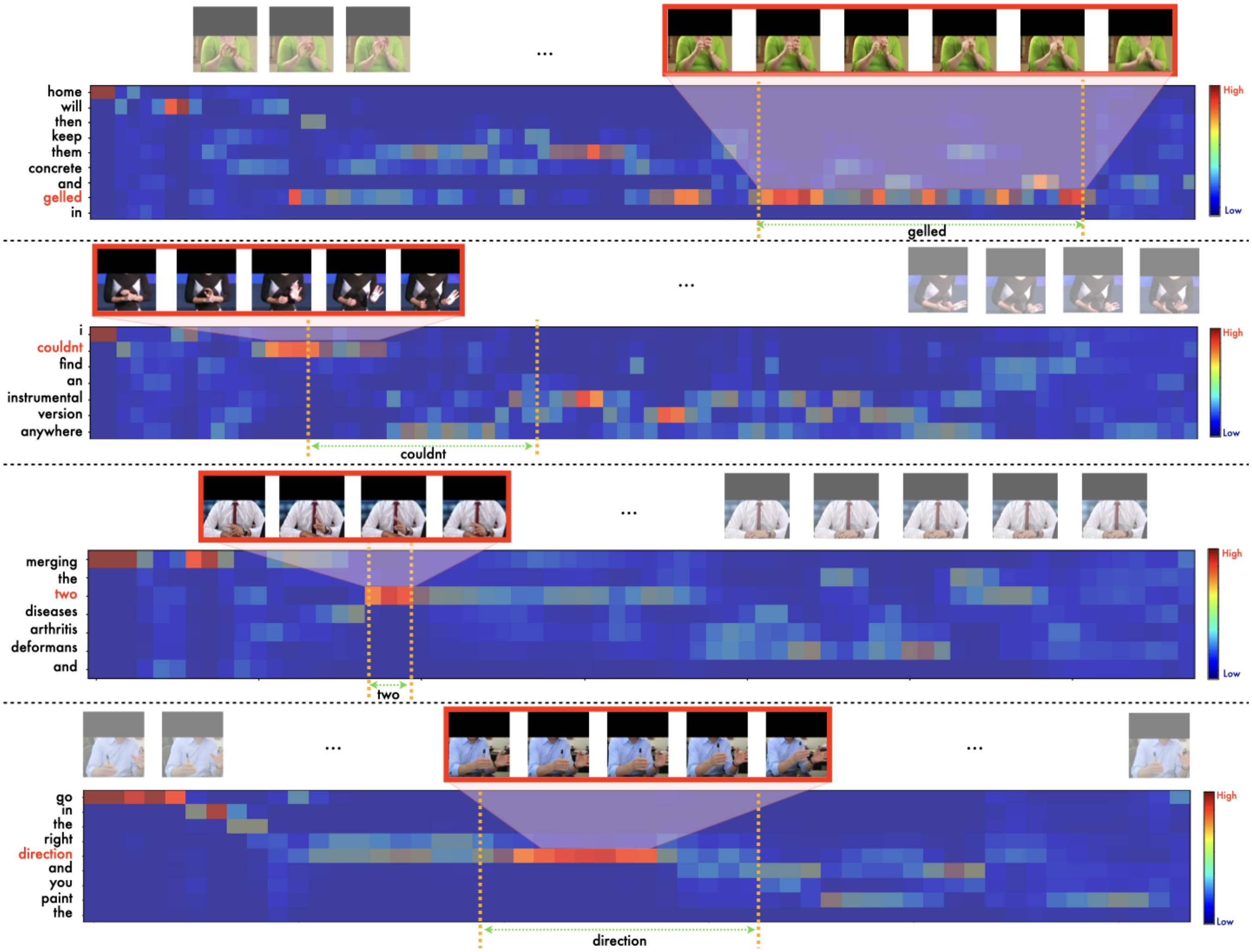

Figure 5. Additional gestured word spotting results on AVS-Spot dataset. Keywords are highlighted in red on the left panel and the speech-based force alignment word boundaries are marked by yellow lines. JEGAL successfully spots the gestured keywords, demonstrating its robustness across diverse gestures and speakers. The red triangles zoom into the corresponding frames where JEGAL detects the keywords, clearly aligning with the gestures. Note that in some cases (e.g., rows 2 and 4), ground-truth boundaries may slightly differ, as the speaker can gesture and utter the same word at slightly different times. JEGAL effectively estimates the approximate intervals where the target word is gestured.

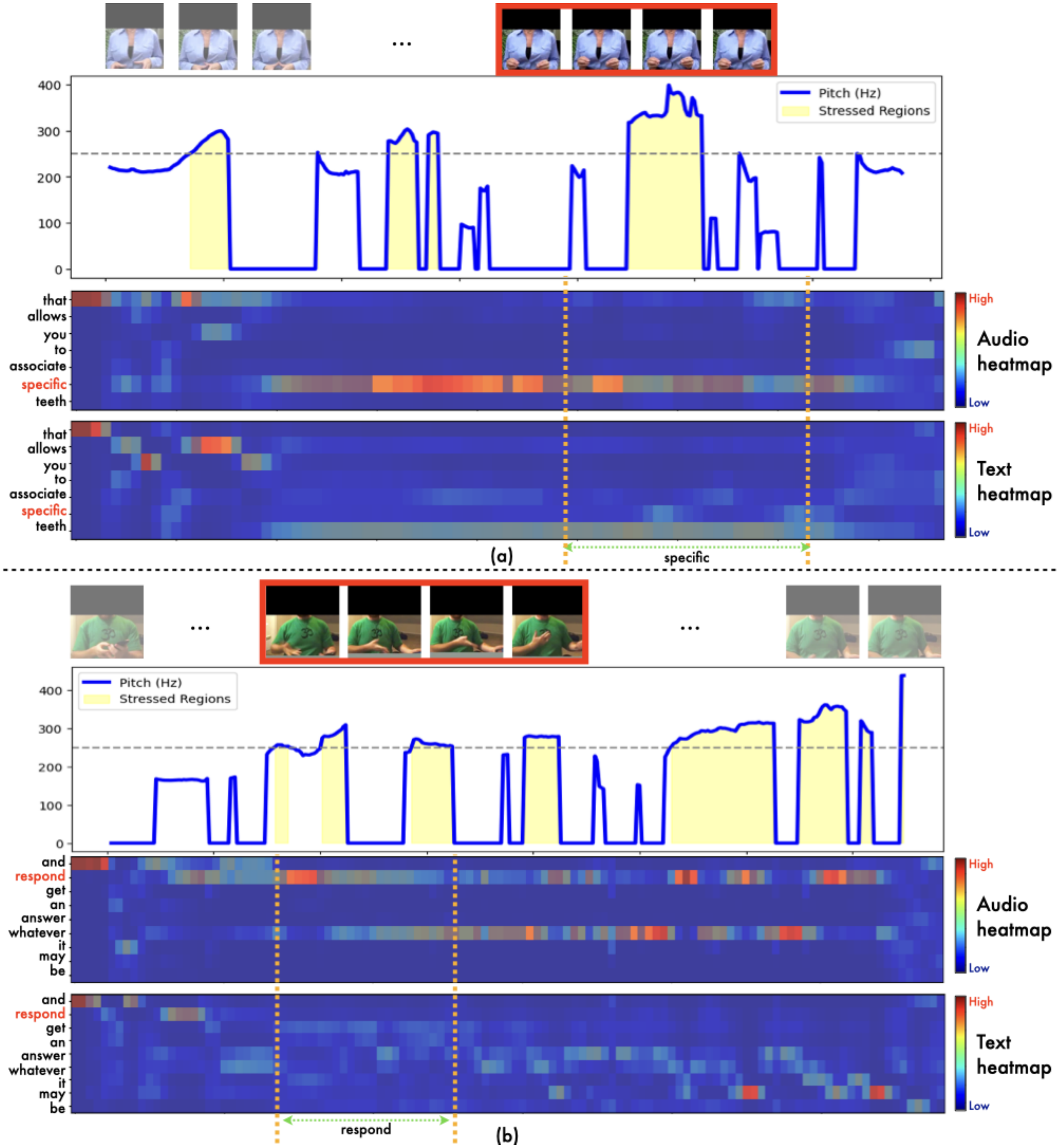

Figure 6. Examples highlighting the role of stressed speech regions in audio-based gesture spotting. The audio-only model successfully detects the stressed keywords "specific" and "respond", whereas the text-only model misses these words. Evidently, the audio-only model looks for word emphasis cues (indicated by high pitch) as such words are more likely to be gestured. This would be difficult to infer from text modality alone. These examples illustrate the advantages of leveraging audio cues for gesture spotting.