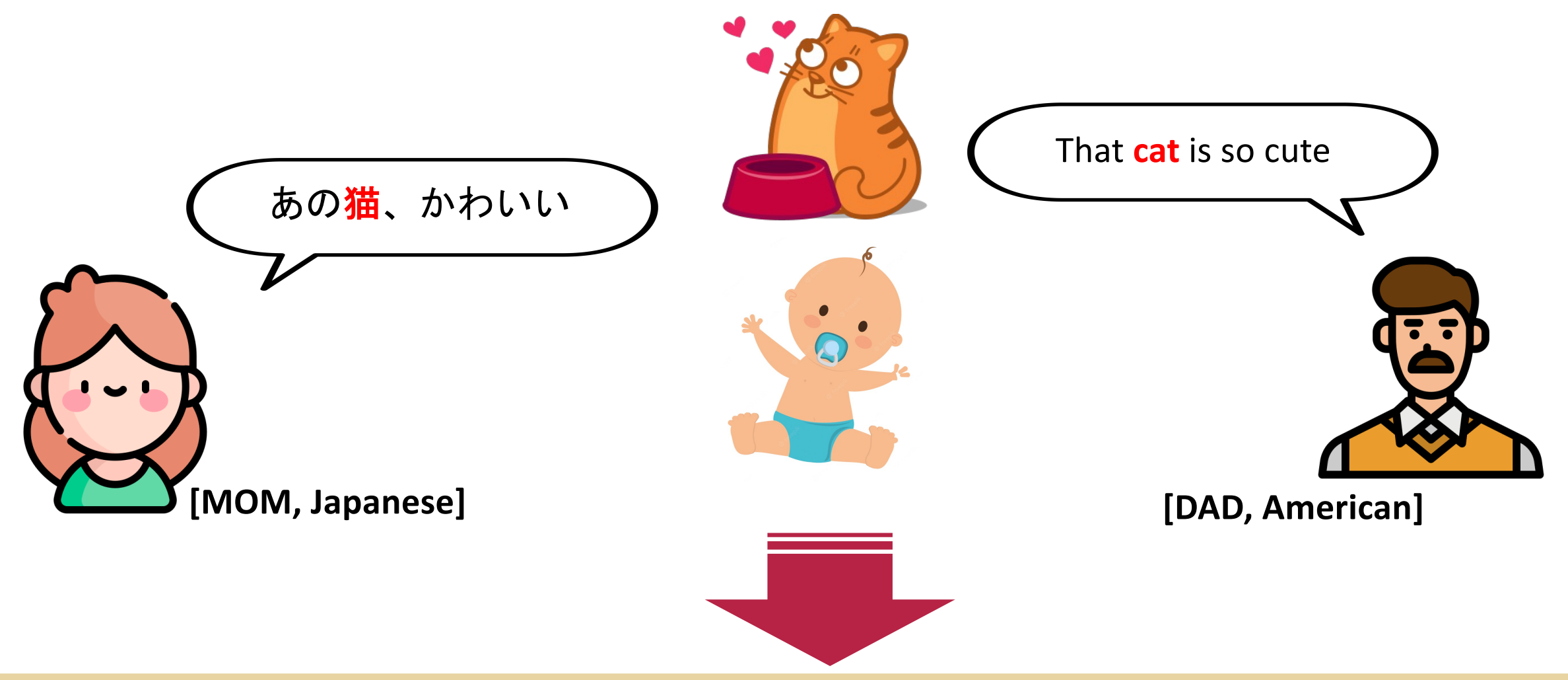
VGSAlign: Bilingual Speech Alignment of Unpaired and Untranscribed Languages using Self-Supervised Visually Grounded Speech Models

Luan Thanh Nguyen , Sakriani Sakti @ HA3CI Laboratory, JAIST, Japan
{luannt, ssakti [at] jaist.ac.jp}

INTRODUCTION

- **Speech-to-Speech Translation (S2ST)**
  - Enhance multilingual communication
  - Relies on a massive amount of parallel source-target speech data
  - Parallel data is often unavailable
- **Human Infants**
  - Multilingual acquisition ability
  - Allow them to acquire languages based on visual information



The paper proposes VGSAlign:

- ✓ Attempt to mimic human infants' behavior
- ✓ Aim to discover the speech pairs data for S2ST
- ✓ Find speech similarity of source and target languages based on corresponding visual context
- ✓ Utilize self-supervised visually grounded speech model
- ✓ Unable to deal with S2ST for unknown, unpaired, untranscribed languages

VGSALIGN FRAMEWORK

The system combines two modules:

(1) Image-Image Similarity Module

(2) Cross Speech-Image Similarity Module

Leveraging self-supervised visually grounded speech models as encoders for image and audio

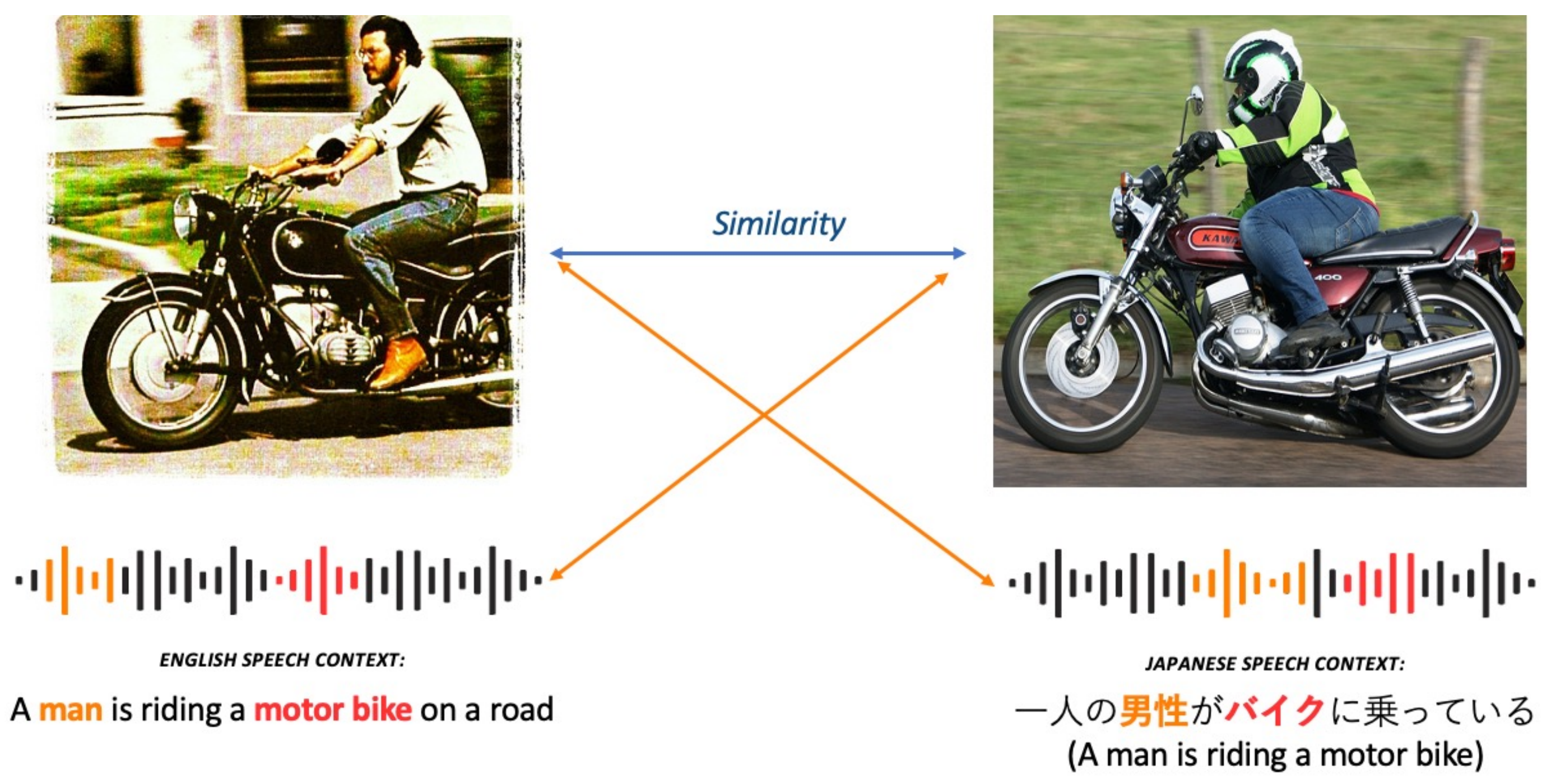

Similarity
ENGLISH SPEECH CONTEXT:
A man is riding a motor bike on a road
JAPANESE SPEECH CONTEXT:
一人の男性がバイクに乗っている
(A man is riding a motor bike)

Figure 1: *Bilingual speech alignment by visual-based information.*
(Note: The image is from the MS-COCO dataset)

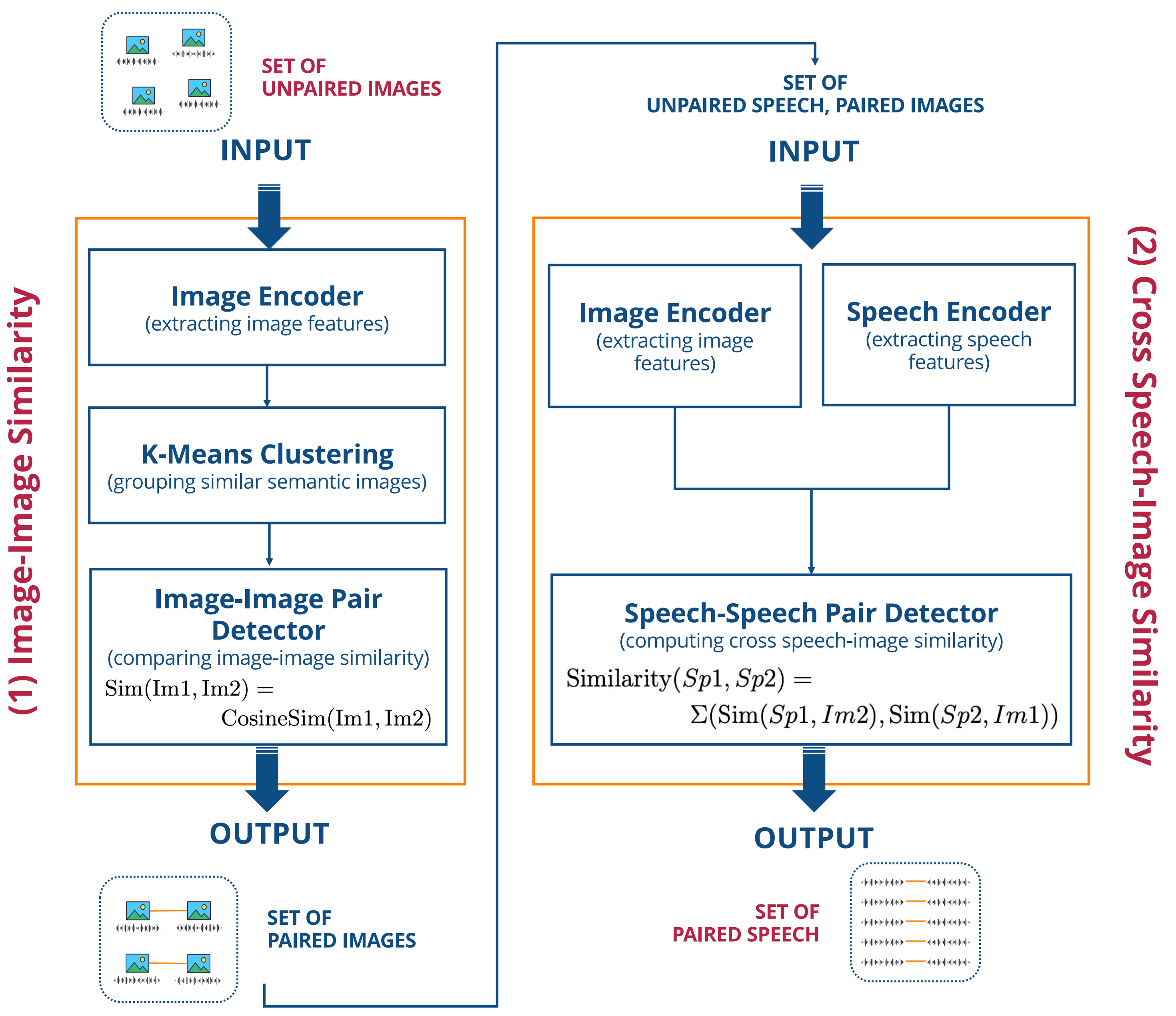

SET OF UNPAIRED IMAGES
INPUT
(1) Image-Image Similarity
Image Encoder
(extracting image features)
K-Means Clustering
(grouping similar semantic images)
Image-Image Pair Detector
(comparing image-image similarity)
$\mathrm{Sim}(\mathrm{Im1}, \mathrm{Im2}) = \mathrm{CosineSim}(\mathrm{Im1}, \mathrm{Im2})$
OUTPUT
SET OF PAIRED IMAGES
SET OF UNPAIRED SPEECH, PAIRED IMAGES
INPUT
(2) Cross Speech-Image Similarity
Image Encoder
(extracting image features)
Speech Encoder
(extracting speech features)
Speech-Speech Pair Detector
(computing cross speech-image similarity)
$\mathrm{Similarity}(Sp1, Sp2) = \Sigma(\mathrm{Sim}(Sp1, Im2), \mathrm{Sim}(Sp2, Im1))$
OUTPUT
SET OF PAIRED SPEECH

Figure 2: *The overview of the VGSAlign framework.*

EXPERIMENTAL SETTINGS & RESULTS

- **Data**
  - SpokenCOCO with around 600K human recordings in English
  - SpokenSTAIR with around 600K synthesized speech in Japanese
- **Self-supervised VGS Models**
  - Speech encoder: HuBERT (base), or Wav2Vec2.0 (base)
  - Image encoder: DINO-ViT small 8x8
- **Training objective with InfoNCE Loss**

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$$

- **Results**
  - The retrieval recall scores for the models (on SpokenCOCO (English) and SpokenSTAIR (Japanese) test sets)

Model		Image → Speech			Speech → Image			Average Speech ↔ Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SpokenCOCO	VG-HuBERT [1]	42.8	73.6	83.9	30.6	60.8	72.8	36.7	67.2	78.4
	EN-VG-W2V2	41.3	72.3	83.8	29.8	60.0	72.8	35.6	67.2	78.4
	EN-VG-HuBERT	44.1	74.2	84.4	31.0	60.6	72.5	**37.6**	**67.4**	**78.5**
SpokenSTAIR	JA-VG-HuBERT	40.3	72.3	83.2	29.7	60.0	72.3	35.0	66.2	77.8
	JA-VG-W2V2	42.0	73.1	83.3	30.3	60.3	72.7	**36.2**	**66.7**	**78.0**

  - The performance of the VGSAlign system on the speech-speech alignment F1 score (%) (determining speech pair ability)

Model	F1-score
Text-text Alignment	84.49
VGSAlign (Speech-speech Alignment)	54.11

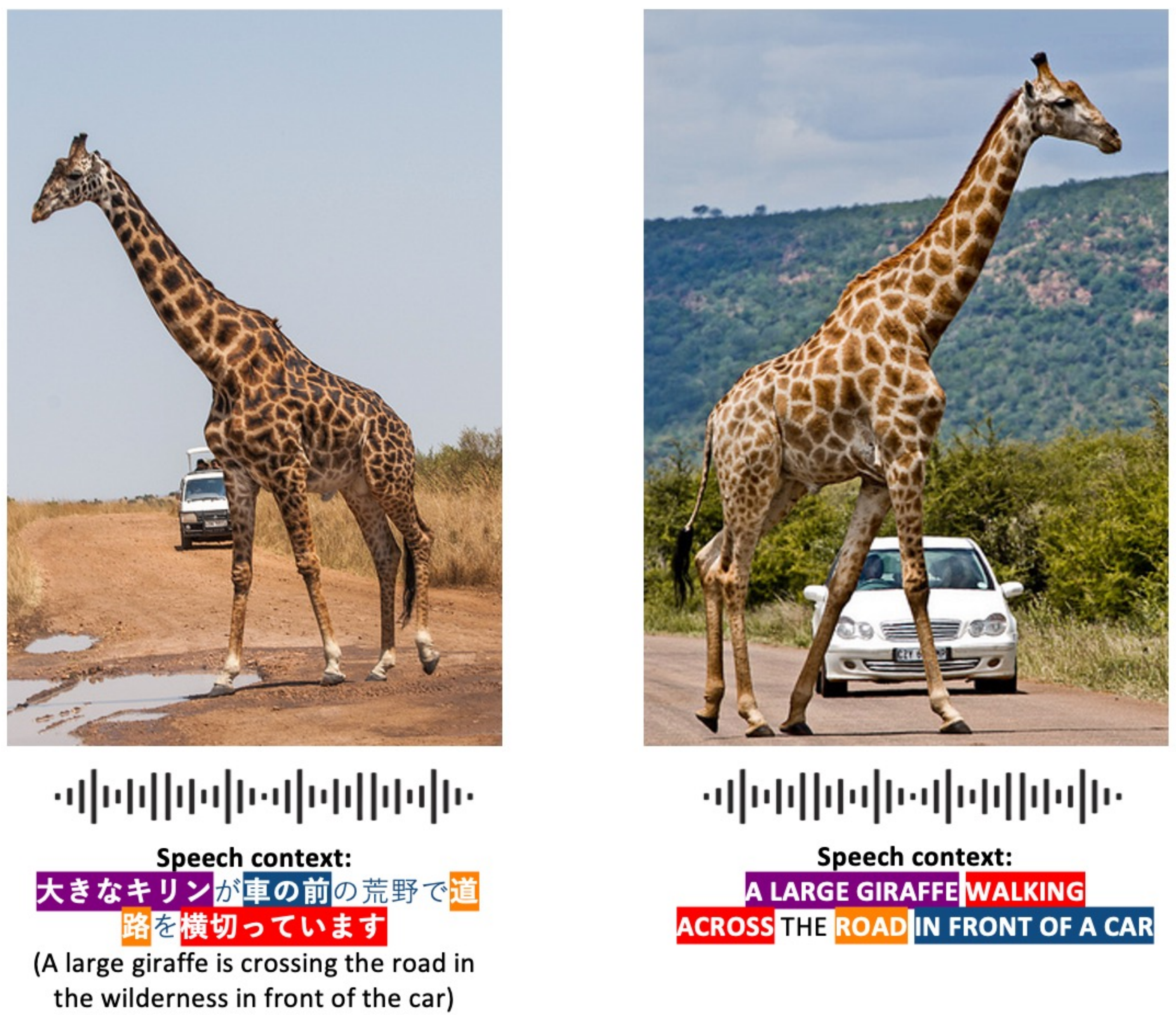

Speech context:
大きなキリンが車の前の荒野で道路を横切っています
(A large giraffe is crossing the road in the wilderness in front of the car)
Speech context:
A LARGE GIRAFFE WALKING ACROSS THE ROAD IN FRONT OF A CAR

Figure 3: *Speech-speech pair determined by the VGSAlign framework.*
(Note: The image is from the MS-COCO dataset)

CONCLUSIONS

- VGSAlign can be applied to any other languages
- Allow mapping speech from the source to the target languages
- Able to determine whether two given speech in different languages are semantically paired without the need for text and knowledge about the language

FUTURE DIRECTIONS

- Perform speech-to-speech translation for unknown, unpaired, untranscribed languages by using data from the VGSAlign system
- Investigate the obtained speech-image co-embeddings in order to get pseudo-speech-speech pairs

REFERENCES

[1] P. Peng et al., "Word discovery in visually grounded, self-supervised speech models", INTERSPEECH 2022

[2] A. Baevski et al., "Wav2Vec 2.0: A Framework for Self-supervised Learning of Speech Representations", NeurIPS 2020

[3] W. Hsu et al., "HuBERT: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units", IEEE/ACM Transactions on Audio, Speech, and Language Processing 2021

In the Proceedings of the 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages, INTERSPEECH Satellite Workshop 2023