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

Luan Thanh Nguyen, Sakriani Sakti

Japan Advanced Institute of Science and Technology, Japan

{luannt,ssakti}@jaist.ac.jp

Abstract

Direct neural speech-to-speech translation (S2ST) systems enable translating speech from source to target languages without the need for text transcription. However, these systems are mostly trained using supervised learning that relies on a massive amount of parallel source-target speech data, which is often unavailable. This paper proposes a bilingual speech alignment approach called VGSAlign, as the initial solution for obtaining paired data from unknown, untranscribed, and unpaired speech data. Here, we assume the speech has auxiliary input from the visual modality that describes the semantic information. The approach then leverages the ability (1) to discover spoken words in multiple languages from the correspondences between speech segments and part of images based on self-supervised visually grounded speech models and (2) to find the visually grounded semantically equivalent between the spoken discovery of speech segments of source and target languages. By learning the representations of speech and images, VGSAlign shows the potential to achieve bilingual speech alignment based on visual representation. Furthermore, experimental results show that the proposed approach could work effectively with unknown, untranscribed, and unpaired speech without being trained on any supervised tasks.

Index Terms: bilingual speech alignment, self-supervised speech representation, visually-grounded speech model

1. Introduction

Machine translation is a technology in the field of artificial intelligence with the purpose of facilitating communication between people who speak different languages. However, this technology has been primarily focused on written languages [1]. Even for developing speech-to-speech translation (S2ST), the traditional approach uses a cascade manner that concatenates automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS), in which written form as an intermediate modality between these systems is critical [2, 3]. Nevertheless, there are approximately 7,000+ spoken languages, half of which do not have a writing system. Therefore, the trend in recent technologies focused more on developing a direct approach of S2ST based on end-to-end deep learning, enabling people who speak any language to translate their speech without the need for text transcription [4, 5, 6]. The success of this technology will allow individuals from all over the world to communicate more fluidly with one another, regardless of their native tongue.

However, despite the benefits of direct neural S2ST technology, it is essential to note that these systems are primarily trained using supervised learning that relies on a massive amount of parallel speech of both source and target languages. Unfortunately, such datasets are often unavailable. Therefore, much work must be done to ensure that it is accessible and effective for all languages and cultures. Several studies attempted to construct unsupervised S2ST (US2ST) systems. To date, Wang et al. [7] proposed to develop US2ST by cascading unsupervised ASR (UASR) [8], unsupervised machine translation (UMT) [9, 10], and unsupervised TTS (UTTS) [11, 12]. UASR was trained to output pseudo labels given only speech data, and UMT was trained to map source-target monolingual corpus into shared latent representation via adversarial learning. In contrast, UTTS was trained to generate speech waveform given the pseudo labels. However, similar to traditional cascade S2ST, such an approach often suffers from severe error propagation. A study by Fu et al. proposed denoising back-translation to address error propagation problems [13].

In contrast with existing cascade US2ST, we aim at US2ST based on the neural direct translation framework, and this paper proposes a bilingual speech alignment approach called VGSAlign, as the initial solution for obtaining paired source-target speech datasets from unknown, untranscribed, and unpaired monolingual speech data from two different languages. Here, we assume the speech has auxiliary input from the visual modality that describes the semantic information. The approach then leverages the ability (1) to discover spoken words in multiple languages from the correspondences between speech segments and part of images based on self-supervised visually grounded speech models and (2) to find the visually grounded semantically equivalent between the spoken discovery of speech segments of source and target languages. By learning the representations of speech and images, VGSAlign shows the potential to achieve bilingual speech alignment based on visual representation.

2. Related Works

Visual and spoken language representation. Associating the semantic aspect of spoken language representation with images has grown much attention in recent years. Most approaches learn to align visual images and untranscribed spoken captions by modeling image and speech representation in a joint embedding space [15, 16, 17, 18]. On the other hand, research on a self-supervised approach to speech has also become popular as it allows producing acoustic embeddings from the input speech waveform without any supervision. Such an approach has been shown to be effective in producing high-quality speech representations. Hsu et al. [19] presented the Hidden Unit-BERT (HuBERT) model for self-supervised speech representation, which performed well on various speech tasks. A study by Peng et al. proposed a method for visually grounded spoken term discovery by utilizing HuBERT to automatically

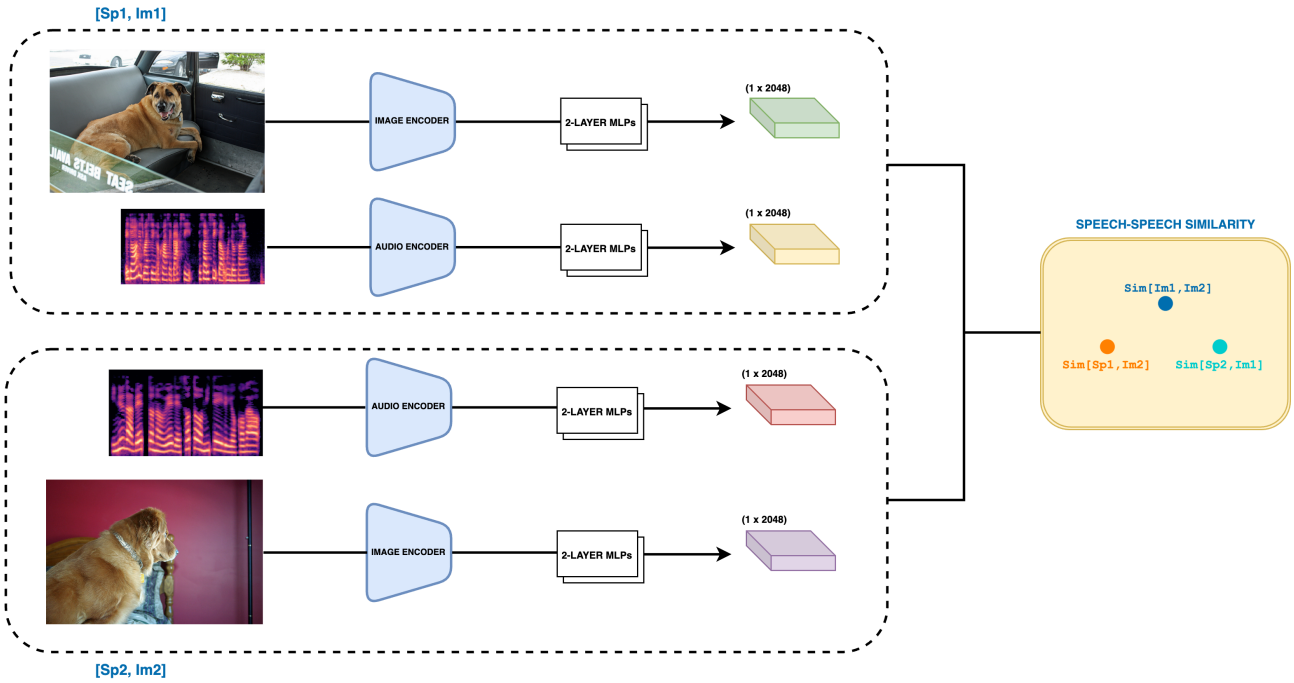


Figure 1: *The overview of the proposed VGSAlign system with the input are speech-image pairs from source and target languages. Note that the images are from the MS-COCO dataset [14].*

discover (localize, segment, and identify) spoken words based on visually grounded models [20]. Unfortunately, these studies mainly focused only on monolingual settings. Several studies then offered to provide multilingual visually-grounded speech models. However, these approaches require equal samples to learn the triple association of one image and two speech representations from two different languages ($Sp1$, Im, $Sp2$) [21]. Ryu et al. investigated the impact when one language has more data than the other to simulate whether richer language can support the under-resourced languages [22]. Nevertheless, the approaches assumed that the images in those languages are the same. Nevertheless, in reality, such conditions are difficult to obtain. In contrast, in our study, we deal with multiple visually grounded speech representations where the images of those languages may be different ($Sp1$, $Im1$, $Im2$, $Sp2$). The idea might be close to Suris et al.'s [23]. However, their approach focused only on discrete transcription, while here, we deal with continuous speech representation without any text information.

Bilingual alignment. The lack of large-scale parallelizing in pairs between source and target data forced many researchers to construct technologies with purely non-parallel data [24, 25]. Wang et al. [26] addressed the problem of non-parallel source and target sentences using partially aligned sentence pairs, which can be incorporated into the conventional training phase of the model. However, this study is originally developed only for text translation. In contrast, our work deals with unknown, untranscribed, and unpaired speech utterances. Furthermore, as the image representations of those bilingual speech segments may be different, we first need to identify the speech pairs ($Sp1$ and $Sp2$) by calculating the similarity between those two image representations ($Im1$ and $Im2$). After that, we discover partially aligned bilingual speech segments on the generated pseudo pair of speech utterances.

3. VGSAlign - Bilingual Speech Alignment

The proposed VGSAlign system aims to achieve the speech pairs between source and target languages based on corresponding visual context. The system combines two self-supervised visually grounded speech models as encoders for image and audio. The datasets for training these models consist of speech Sp and their corresponding images Im. The system is depicted in Figure 1, which shows how the proposed VGSAlign system works to determine whether two speech utterances ($Sp1$ and $Sp2$) are partially semantically paired or not based on the speech-speech similarity (as mentioned in Section 3.2)

3.1. Self-supervised Visually Grounded Speech Model

Our self-supervised VGS models follow the structure of the research of Peng et al. [20]. The model has a dual-encoder architecture, including (1) an audio encoder based on a self-supervised speech model such as HuBERT [19] or Wav2Vec2.0 (W2V2) [3] and (2) an image encoder is a self-supervised vision transformer model as DINO-ViT [27]. Input to the model consists of a raw speech waveform and its corresponding image. After being fed respectively into the audio and image encoders, the output of the self-supervised VGS model is the similarity score between them, indicating the speech reflects the content of the image when it is large. In contrast, the opposite is true for the small similarity score. The model is trained using the InfoNCE loss [28, 29], which is effective for various self-supervised learning tasks.

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k\left(x_{t+k}, c_t\right)}{\sum_{x_j \in X} f_k\left(x_j, c_t\right)} \right] \tag{1}$$

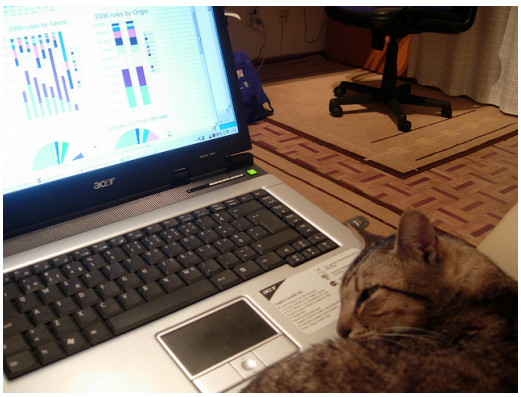
The InfoNCE loss attempt to optimize a given expression by considering a collection $X = \{x_1, \ldots, x_N\}$ comprising N

random samples. This collection includes a single positive sample drawn from the distribution $p(x_{t+k} \mid c_t)$, where c_t represents a specific condition at time $t+k$. Additionally, X contains $N-1$ negative samples obtained from the 'proposal' distribution $p(x_{t+k})$. To this end, this loss aims to maximize the high similarity scores to related speech-image pairs and otherwise.

3.2. Bilingual Speech Alignment

The main scenario in this work is that each speech has its corresponding image, and both images of two speech have common semantic parts, as shown in the example in Figure 2 below.

Speech context: A cat is sitting on a desk eating off a plate



Speech context: 仕事中のパソコンの前を猫が陣取る
(A cat is sitting in front of a desk computer)

Figure 2: *The scenario in this work with two speech in different languages describing two images. Note that the two images are partially semantically paired.*

First, we process the set of images and extract their features by an image encoder. Then, we employ K-means clustering on the extracted image features to group the images into clusters based on their similarities. This clustering approach allows us to gather images with similar characteristics together, thereby reducing the computation space.

At each cluster, we compute the similarity between one image $Im1$ to other images $Im2$ within the cluster. The image with the most similarity $Sim(Im1, Im2)$ is chosen as an initial pair. Note that although $Im1$ and $Im2$ are different, they include some parts with similar semantic information. The remaining clusters also go through the same process.

For each image, we initialy have two corresponding speech, Sp_{source} and Sp_{target}, from two different languages. We only keep the Sp_{source} (or $Sp1$) of $Im1$, and Sp_{target} (or $Sp2$) of $Im2$ for the initial image pair $(Im1, Im2)$. Consequently, the speech $Sp1$ of $Im1$ is temporarily seen as a paired speech of the speech $Sp2$ of $Im2$.

After that, to determine whether those two speech are paired or not, inspired by the work of Suris et al. [23] with the text-text similarity based on visual representation, we consider the speech-speech similarity defined as 2 below.

$$\begin{aligned} &Sim(Sp1, Sp2) \\ &= \Sigma(Sim(Sp1, Im2), Sim(Sp2, Im1)) \end{aligned} \tag{2}$$

To decide a couple of speech, as defined above, is paired or not, we compute the cross similarity $(Sp1, Im2)$ between the source speech and target image, and the same with $(Sp2, Im1)$ by conducting dot product the output of the VGS model encoders. Both image and speech are fed into respective encoders of the self-supervised VGS models. Once the features have been extracted, we use 2-layer MLPs to project them into a 2048-dimensional space. Their final sum similarity between the cross similarity is used to determine whether those two input speech are partially paired or not by using an appropriate threshold.

4. Experiments

4.1. Data

We perform our experiments using the SpokenCOCO [30] and SpokenSTAIR [31] for English and Japanese speech corpora, respectively. SpokenCOCO is a dataset that contains approximately 600K recordings of human speakers reading the MS-COCO [14] image captions out loud in English. For Japanese speech, SpokenSTAIR is generated by synthesizing captions from the STAIR dataset [32] following the same methodology as Chrupala et al. [16]. The SpokenSTAIR also consists of around 600K synthesized audio, and there are five spoken captions for each image as the structure of the SpokenCOCO dataset. Note that the images in the SpokenSTAIR dataset are also sourced from the MS-COCO dataset. The proportion of data follows Karpathy split [33] with around 550K, 25K, and 25K audio files for training, validation, and test set, respectively, for both SpokenCOCO and SpokenSTAIR.

4.2. System Setup

We train the self-supervised VGS models with the settings as those of the work of Peng et al. [20]. The audio encoder we use is HuBERT or W2V2, while we use the DINO-ViT as the image encoder. Before calculating the dot product between the outputs of audio and image encoders, we individually transform them using 2-layer MLPs, projecting them into a 2048-dim space. Furthermore, we increase the batch size to 128 and train the self-supervised VGS models over 30 epochs. Models on SpokenCOCO and SpokenSTAIR are trained on a single NVIDIA A6000 GPU for approximately three days.

Table 1: *The retrieval recall scores for the models on SpokenCOCO (English) and SpokenSTAIR (Japanese) test sets, respectively.*

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 [20]	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**

First, we construct a subsystem to choose the initial image pairs for bilingual alignment. We conduct experiments with 1,000 images randomly taken from the test set of MS-COCO. As a result, the number of obtained image-image, as well as speech-speech pairs, is 500. Note that each image has two corresponding speech from SpokenCOCO and SpokenSTAIR, and both describe the same image in English and Japanese. The VGG-16 [34] is used as an image feature extractor. We then classify images using K-means clustering with K equal to 10. To calculate the similarity between an image and the remaining images in each cluster, we use Cosine Similarity.

After obtaining the initial image pairs, for each pair $(Im1, Im2)$, we keep only the Japanese speech for $Im1$ and only English speech for $Im2$. The English and Japanese speech are fed into the respective encoders, and the same with their corresponding images. The extracted features of the two speech-image pairs are then used for calculating the speech-speech similarity (as mentioned in Equation 2). We scale the range of similarity from [-1, 1] up to [0, 1] and set a threshold as 0.5 to decide whether the two input speech are a pair.

4.3. Evaluation Metrics and Results

Speech-Image Retrieval Recall Score. As multimodal models, we first evaluate all self-supervised VGS models on their retrieval performance by retrieval recall (R@K) score. The R@K with the K at 1, 5, and 10 are depicted in Table 1. With the obtained results, we can see that on the average speech-image retrieval recall score, our re-trained EN-VG-HuBERT slightly outperforms the VG-HuBERT on the SpokenCOCO and JA-VG-W2V2 achieves the highest results on the SpokenSTAIR. Hence, EN-VG-HuBERT and JA-VG-W2V2 are chosen as the self-supervised VGS models for the VGSAlign system.

Speech-speech Alignment Score. We compute the F1 score towards the proposed speech-speech similarity described in Equation 2 to evaluate the speech-speech alignment performance. For comparison, we also calculate topline text-text similarity when transcription is known. Here, the text-text similarity is determined by computing cosine similarity [35] between the sentence embeddings extracted from text captions of the spoken speech captions. The obtained results are shown in Table 2 below. Note that we evaluate the part of 1,000 images from the test set of the SpokenCOCO and SpokenSTAIR datasets, the same part as the one used for experiments in Section 4.2.

Table 2: *The performance of the VGSAlign system on the speech-speech alignment F1 score (%).*

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

From the result of the VGSAlign performance, we can see that although the speech-speech alignment result is lower than the text-text alignment, it shows that this approach can work with the scenario of unpaired and untranscribed languages for deciding speech pairs based on visual semantic information.

4.4. Discussion

Speech-image alignment ability of self-supervised VGS models. From the initial alignment results on the SpokenCOCO and SpokenSTAIR dataset to find the pair without the demand of texts, we hope that the VGSAlign works well on other unpaired and untranscribed languages thanks to the ability of self-supervised learning models to learn the speech presentation from unlabeled data. As shown in Table 1, the self-supervised VGS models show their ability to learn the co-representation to find the similarity between speech and its image, which is crucial for aligning two speech from different languages.

Speech-speech alignment ability. With the achieved results, it shows that the VGSAlign system has the potential to determine whether two speech samples from different languages are paired or not, even without any demand for text. This is a significant premise, particularly since the lack of paired speech data often hinders the task of direct neural speech-to-speech translation. Training the models not on any supervised tasks, the proposed VGSAlign system is a solution to process unknown, untranscribed, and unpaired speech for determining and choosing data for direct neural speech-to-speech translation.

5. Conclusion

In conclusion, our study demonstrated a proposed speech-speech alignment called VGSAlign based on self-supervised VGS models to find the similarity between speech from source and target languages. We verified that without the need for text and knowledge about the language, the system could determine whether two given speeches in different languages are semantically paired by computing their similarity.

In future, we plan to do experiments and assess the performance of speech-to-speech translation using data from the VGSAlign system. Moreover, for the speech-image alignment, we intend to investigate the speech-image co-embedding obtained by the speech and image encoders of the self-supervised VGS models to get partial pseudo-pairs of speech and image.

6. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP21H03467.

7. References

[1] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, 2022.

[2] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The atr multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 365 – 376, 04 2006.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *ArXiv*, vol. abs/1904.06037, 2019.

[5] T. Kano, S. Sakti, and S. Nakamura, "End-to-end speech translation with transcoding by multi-task learning for distant language pairs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1342–1355, 2020.

[6] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, 2022, pp. 860–872.

[7] C. Wang, H. Inaguma, P.-J. Chen, I. Kulikov, Y. Tang, W.-N. Hsu, M. Auli, and J. Pino, "Simple and effective unsupervised speech translation," *ArXiv*, vol. abs/2210.10191, 2022.

[8] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Advances in Neural Information Processing Systems*, 2021, pp. 27 826–27 839.

[9] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," *ArXiv*, vol. abs/1711.00043, 2017.

[10] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *ArXiv*, vol. abs/1901.07291, 2019.

[11] J. Ni, L. Wang, H. Gao, K. Qian, Y. Zhang, S. Chang, and M. H. Johnson, "Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition," *ArXiv*, vol. abs/2203.15796, 2022.

[12] A. H. Liu, C.-I. J. Lai, W.-N. Hsu, M. Auli, A. Baevskiv, and J. Glass, "Simple and effective unsupervised speech synthesis," *ArXiv*, vol. abs/2204.02524, 2022.

[13] Y.-K. Fu, L.-H. Tseng, J. Shi, C.-A. Li, T.-Y. Hsu, S. Watanabe, and H. yi Lee, "Improving cascaded unsupervised speech translation with denoising back-translation," *ArXiv*, vol. abs/2305.07455, 2017.

[14] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Doll'a r, and C. L. Zitnick, "Microsoft COCO: common objects in context," *ArXiv*, vol. abs/1405.0312, 2014.

[15] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016.

[16] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 613–622.

[17] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 506–517.

[18] H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," *arXiv preprint arXiv:1703.08136*, 2017.

[19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[20] P. Peng and D. Harwath, "Word discovery in visually grounded, self-supervised speech models," in *Proc. INTERSPEECH 2022*, 2022, pp. 2823–2827.

[21] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4969–4973.

[22] H. Ryu, A. Senocak, I. S. Kweon, and J. S. Chung, "Hindi as a second language: Improving visually grounded speech with semantically similar samples," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[23] D. Surís, D. Epstein, and C. Vondrick, "Globetrotter: Connecting languages by connecting images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 474–16 484.

[24] D. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, pp. 477–504, 12 2005.

[25] L. Chen, "Machine translation model based on non-parallel corpus and semi-supervised transductive learning," *ArXiv*, vol. abs/1405.5654, 2014.

[26] Y. Wang, Y. Zhao, J. Zhang, C. Zong, and Z. Xue, "Towards neural machine translation with partially aligned corpora," in *International Joint Conference on Natural Language Processing*, 2017.

[27] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[28] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018.

[29] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China, 2019, pp. 55–65.

[30] W.-N. Hsu, D. Harwath, T. Miller, C. Song, and J. Glass, "Text-free image-to-speech synthesis using learned segmental units," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5284–5300.

[31] W. N. Havard, J. Chevrot, and L. Besacier, "Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8618–8622.

[32] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, 2017, pp. 417–421.

[33] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020.