Multilingual Self-Supervised Visually Grounded Speech Models

Huynh Phuong Thanh Nguyen[1], Sakriani Sakti[1,2]
[1]Japan Advanced Institute of Science and Technology, Japan
[2]Nara Institute of Science and Technology, Japan
{s2210406,ssakti}@jaist.ac.jp

Abstract

Developing a multilingual speech-to-speech translation system poses challenges due to the scarcity of paired speech data in various languages, particularly when dealing with unknown and untranscribed languages. However, the shared semantic representation across multiple languages presents an opportunity to build a translation system based on images. Recently, researchers have explored methods for aligning bilingual speech as a novel approach to discovering speech pairs using semantic images from unknown and untranscribed speech. These aligned speech pairs can then be utilized to train speech-to-speech translation systems. Our research builds upon these approaches by expanding into multiple languages and focusing on achieving multimodal multilingual pairs alignment, with a key component being multilingual visually grounded speech models. The objectives of our research are twofold: (1) to create visually grounded speech datasets for English, Japanese, Indonesian, and Vietnamese, and (2) to develop self-supervised visually grounded speech models for these languages. Our experiments have demonstrated the feasibility of this approach, showcasing the ability to retrieve associations between speeches and images. The results indicate that our multilingual visually grounded speech models yield promising outcomes in representing speeches using semantic images across multiple languages.

Keywords: multilingual visually grounded speech models, self-supervised speech representation, speech translation

1. Introduction

Speech translation is important in bridging the communication gap between individuals who speak different languages. There are various methods proposed for enabling communication across diverse languages, such as speech-to-speech translation (S2ST) (Nakamura, 2009; Shimizu et al., 2008). Additionally, text-less S2ST systems have also been developed using end-to-end deep learning (Li et al., 2023; Lee et al., 2022). However, these techniques pose significant challenges that need to be overcome, such as the lack of parallel source-target data or unbalanced data between two languages.

The fact that multiple languages can share the same semantic image presents an opportunity to develop a multilingual speech-to-speech translation system based on images. Recently, bilingual speech alignment methods which involve matching spoken words or sounds in one language with their corresponding counterparts in another language, have been explored as a novel approach to translate speech between two languages using semantic images. VGSAlign has been introduced (Nguyen and Sakti, 2023) as an example. It involves using speech alignment of unpaired and untranscribed data. Self-supervised Visually Grounded Speech (VGS) model is a model that integrates visual information such as images with speech signals to perform speech-related tasks. It is used to find visually grounded semantically equivalent parts between the speech segments of the source and target languages. According to the results from VGSAlign research, this approach shows potential applicability in bilingual speech alignment without being trained on any supervised tasks.

Taking inspiration from the VGSAlign framework, our goal is to achieve multilingual self-supervised VGS models as an extension of the VGSAlign framework. These models can be used to extract semantic information for multilingual speech alignment. We have specifically selected English (VN), Japanese (JA), Indonesian (ID), and Vietnamese (VN) as the target languages. The main contributions of this research are to (1) generate VGS datasets for four languages using text-to-speech synthesis as the core technique and (2) achieve the multilingual self-supervised VGS models through fine-tuning and further training strategies based on the VGSAlign framework.

2. Related Works

In recent years, the use of visually grounded models has become a popular method among researchers to address issues of speech and text alignment. These techniques employ visual presentation to align different items with the same meaning. Additionally, the visually grounded models also contribute to the reduction of resource challenges. Given the fact that acquiring image datasets is relatively easier due to the huge amount of available resources and the ease of generating them. A method was proposed for visually grounded spoken term discovery, which aims to associate spoken captions with natural images (Peng and Har-

SIGUL2024 Workshop, pages 237–243
21-22 May, 2024.

wath, 2022). This resulted in the automatic discovery of words in a speech signal, including localization, segmentation, and identification. The results suggest that a computational model can learn the structure of spoken language from untranscribed speech audio using a combination of multiple self-supervised objectives. Unfortunately, these studies mainly focused only on monolingual settings.

Furthermore, the paper (Kamper and Roth, 2018) demonstrates the ability to apply the visual grounding in cross-lingual keywords, yielding high retrieval results. Other approaches used a joint embedding space for modeling image and speech representations to align visual images with untranscribed spoken captions(Harwath et al., 2016; Harwath and Glass, 2017; Kamper et al., 2017). Chrupała et al. presented a visually grounded model of speech perception that projects speeches and images into a joint semantic space (Chrupała et al., 2017). This research demonstrates the potential of the visual grounding method, which extracts semantic information from images to align both speech and text.

Several studies have proposed models for multilingual visually grounded speech. These models, however, require balanced datasets to learn the triple association between an image and two speech representations from different languages *(Sp1, Im, Sp2)* (Harwath et al., 2018). Ryu explored the effect of language data imbalance. This paper stated that in a bilingual VGS model, a high-resource language can enhance the performance of a low-resource language by using semantically similar spoken captions. (Ryu et al., 2023). These studies also assumed identical images or captions across languages, which is not available. VGSAlign offers a solution for handling multiple visually grounded speech representations where the images in different languages may not be the same *(Sp1, Im1, Im2, Sp2)*. It also handles continuous speech representation without relying on any text information, successfully achieving bilingual speech alignment for unpaired and untranscribed languages.

Our ongoing research aims to extend VGSAlign to accommodate multilingual speech alignment, with a focus on four languages: English, Japanese, Indonesian, and Vietnamese.

3. System

3.1. Multilingual VGS Model

The objective of our research is to achieve the multilingual self-supervised Visually Grounded Speech Model (VGS Model), which serves as an extension of the self-supervised VGS model in the VGSAlign framework proposed in the paper (Nguyen and Sakti, 2023). Expanding on the based model from VGSAlign, our research makes contributions by continuing to train this model using data from the Flickr8K dataset for four different languages (*EN*, *JA*, *ID*, *VN*). The training datasets for these models consist of pairs of speech *Sp* and corresponding image *Im*.

3.2. VGSAlign-Based Framework

The VGSAlign (Bilingual Speech Alignment) framework aims to align speech between source and target languages based on corresponding visual context. This system combines two self-supervised models grounded in visual information, serving as encoders for images and audio.

The structure of the self-supervised VGS model within the VGSAlign framework is responsible for extracting features and is used for speech alignment in the next stage. According to the figure, the model features a dual-encoder architecture, comprising (1) an audio encoder based on a self-supervised speech model such as HuBERT (Hsu et al., 2021) or W2V2 (Baevski et al., 2020), and (2) an image encoder using a self-supervised vision transformer model like DINO-ViT (Caron et al., 2021). Then, both audio and image encoders are individually transformed using 2-layer MLPs, projecting them into a 2048-dim space. A pair of images and their corresponding audio are used as input to the model. The output of the self-supervised VGS model is a similarity score indicating how well the speech reflects the content of the image. The InfoNCE loss (Oord et al., 2018; Ilharco et al., 2019) is used to maximize the similarity scores for related speech-image pairs in the training procedure.

3.3. Fine-Tuning and Further Training Strategies for Self-Supervised Visually Grounded Speech Model

The multilingual self-supervised VGS Model is achieved by utilizing a training strategy that uses fine-tuning and further training on the based VGS model. Figure 1 visualizes the process of generating the multilingual VGS model based on the based models in VGSAlign. The EN and JA pre-trained VGS models are used to fine-tune using the EN and JA VGS datasets. The best checkpoints from the pre-trained models are resumed, which retains all the training parameters from the previous research, allowing us to continue fine-tuning the model with new datasets. Moreover, due to the lack of language-compatible VGS available models for ID and VN datasets, the EN pre-trained model is used to continue learning with these datasets.

Additionally, for ID and VN datasets, due to the lack of the language-compatible VGSAlign available models, the EN pre-trained model is used as a standard model to continue learning with

ID and VN datasets. Based on the results of VGSAlign, the model trained with the EN Spoken-COCO dataset achieved better performance compared to the model trained with the JA Spoken-STAIR dataset. This motivated us to use the EN pre-trained model as the base model for training with VN and ID datasets.

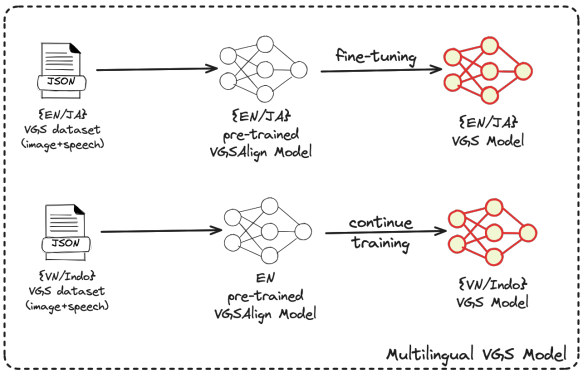

Figure 1: The overview of the fine-tuning and further training strategies for the Multilingual VGS Model.

After the training procedure, the multilingual VGS Model is obtained, which includes two self-supervised VGS models: (1) EN-VGS-Model trained with three languages: EN, ID, and VN, and (2) JA-VGS-Model trained with one language: JA.

4. Experiments

4.1. Data Preparation

This research uses the Flickr8K (Harwath and Glass, 2015) as the main dataset for improving and testing the models. The data proportions follow the original Flickr8K split (Herman Kamper, Mark Hasegawa-Johnson, 2018), with 6K, 1K, and 1K data allocated for the training, validation, and test sets, respectively. To enhance the model with multilingual capabilities, datasets in four languages are used. The data structure follows the structure described in the paper (Harwath and Glass, 2015), which contains pairs of images and their corresponding speech. However, the lack of datasets in JP, ID, and VN posed challenges in collecting complete datasets for the learning process. As a solution, we generate datasets for all three languages based on the English dataset.

4.1.1. Data Generation

Figure 2 visualizes the process of generating datasets for the JA, ID, and VN languages. In this process, the caption datasets in three languages, obtained from (Herman Kamper, Mark Hasegawa-Johnson, 2018; Nugraha et al., 2019; Pham Thanh Trung, 2022), are used as the textual input for our data synthesis pipeline. We use the Text-to-Speech (TTS) synthesis model available from the Google API (Google), specifically the WaveNet architecture, to convert the textual captions into speech audio. The WaveNet model is well-known for its ability to generate highly natural-sounding speech, which is crucial for maintaining the quality and authenticity of the synthesized datasets. The speech synthesis follows a 16kHz and MP3 audio structure, as described in (Nguyen and Sakti, 2023) paper. Next, we combine the synthesized speech datasets with image datasets to obtain the VGS datasets that are formatted as JSON files, containing the pairs of image data along with the corresponding synthesized speech audio. By following this process, we generate a collection of multilingual datasets that support this research.

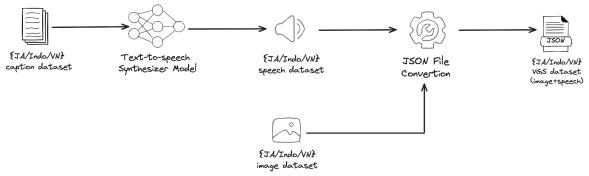

Figure 2: VGS datasets generation process.

4.1.2. Data Analysis

After completing the data generation process, there are a total of four datasets in four languages. Each dataset contains 8000 pairs of images and their corresponding speech that describes the image. The images in each dataset are the same as those in the English dataset, which is considered the standard.

From the initial Flickr8K dataset, there are a total of 8000 images, with each image having 5 different captions. For the English dataset, we choose the first caption for each image and pair it with the corresponding audio. The second, third, and fourth captions belong to the Japanese, Vietnamese, and Indonesian datasets, respectively. As a result, although the four datasets share the same images, the captions differ across languages. This approach ensures a variety of linguistic descriptions for identical sets of images.

4.2. Model Setup

Our self-supervised VGS models are trained using the same basic settings as the base models in VGSAlign. In the pre-trained models, we utilize HuBERT as the audio encoder instead of using both HuBERT and W2V2, while employing DINO-ViT as the image encoder. Additionally, we reduced the validation batch size to 32 as well as the number of epochs to 20, considering that the size of Flickr8K is much smaller than the SpokenCOCO dataset used

Table 1: The retrieval recall scores of the comparison between the based-VGS models and extended-VGS models on the EN, JA, ID, and VN test sets, respectively.

Model/Languages		Image → Speech			Speech → Image			Average Speech ↔ Image		
		R@100	R@10	R@5	R@100	R@10	R@5	R@100	R@10	R@5
Based-VGS-Models	EN-VGS dataset	0.959	0.717	0.587	0.957	0.720	0.595	0.958	0.718	0.591
	JA-VGS dataset	0.614	0.349	0.229	0.616	0.333	0.212	0.615	0.341	0.221
	ID-VGS dataset	0.302	0.234	0.151	0.289	0.266	0.156	0.296	0.250	0.154
	VN-VGS dataset	0.278	0.216	0.140	0.290	0.234	0.180	0.284	0.225	0.160
Extended-VGS-Models	EN-VGS dataset	0.964	0.726	0.595	0.962	0.719	0.606	**0.963**	**0.722**	**0.601**
	JA-VGS dataset	0.888	0.544	0.435	0.889	0.533	0.426	**0.889**	**0.538**	**0.430**
	ID-VGS dataset	0.418	0.333	0.212	0.408	0.354	0.232	**0.414**	**0.344**	**0.222**
	VN-VGS dataset	0.387	0.324	0.220	0.411	0.360	0.240	**0.399**	**0.342**	**0.230**

in the base models. Our VGS models are trained on a single NVIDIA A6000 GPU for approximately 4 days for the entire dataset of four languages.

First, during the training process for each VGS dataset, a total of 6,000 pairs of images and their corresponding speech are used. This training set provides input to the model and enables it to learn and capture the necessary information to improve its performance. Additionally, a separate validation set consisting of 1,000 values is utilized to validate and optimize the model. Adjustments and improvements are made to the learning parameters based on this validation set. Finally, 1,000 values in the test set are used to evaluate the performance of this trained model.

4.3. Evaluation Metrics and Results

The VGS models are evaluated based on their retrieval performance using the **Speech-Image Retrieval Recall Score (R@K)**. Table 1 shows the R@K scores at K values of 5, 10, and 100, measured in the test set before and after training VGS models. In these evaluation metrics, we assess the retrieval performance for both audio-to-image and image-to-audio. We then calculate the average performance for both directions to evaluate the reflection between image and speech.

According to Table 1, the recall scores for speech-image retrieval significantly improved after applying enhanced training strategies to the base models, compared to using the original based models on the Flickr8K. By fine-tuning the models on the EN and JA datasets, the scores improved for both EN the JA dataset. The improvement in the EN dataset was a minority, while it showed a better increase in the JA dataset. This can be explained that the learning parameters of the based pre-trained model are better optimized for the EN dataset, resulting in higher scores compared to the based pre-trained model in the JA dataset. Therefore, by fine-tuning with other datasets, the performance of the JA-VGS model can be greatly enhanced. Additionally, with continued training on the ID and VN datasets, the results also showed slight improvements in all K-values metrics, with around 5%- 10% improvement.

Moreover, similarity scores are calculated to analyze the closeness of the embedding for the multimodal of speech and image, in comparison with multilingual languages (EN, JA, VN, ID). These calculations are based on the same content: four different pictures of a cat, each associated with audio in a different language. Cosine similarity is utilized for this similarity computation. Figure 3 shows the visualization of these similarities using the t-SNE algorithm (Hinton and van der Maaten, 2008) to reduce the size.

Given a pair of an image and its corresponding audio, the model extracted their features using the Image Decoder and Audio Decoder mentioned in Section 3.2. Figure 3 indicates speech and image representation of semantic "cat" in four languages. The figure illustrates the distances between the image and speech of each language as a visualization of the retrieval recall scores listed in Table 1. The distribution of features in images and speech are different. A greater distance reflects low similarity, while a shorter distance indicates high similarity. As outlined in the retrieval recall results across four languages, the model trained with English and Japanese achieves the highest scores. This suggests that the distance between images and speech is close across all items, as represented by the red color. In contrast, the larger distances between the blue and green points indicate lower retrieval recall scores for these languages compared to the English and Japanese-based models. These results provide an intuitive understanding of the correlation between speech and image in our VGS models.

This figure also demonstrates the distance between images representing four languages as well as the speeches. The images between the four languages show close distance as they represent the same object with varying backgrounds. However, despite the closeness between image-image, the image-speech distances vary across the four languages leading to the speech-speech distance also changing slightly depending on the language pair.

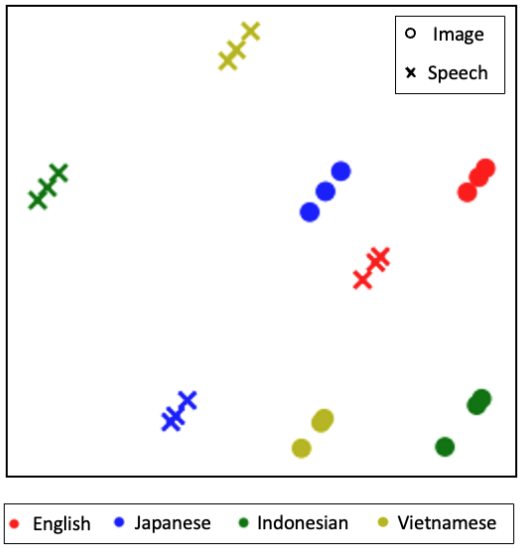

Figure 3: The samples of speech and image representation of the word "Cat" in four languages.

4.4. Discussion

In this paper, we selected Japanese, English, Vietnamese, and Indonesian as a combination of high-resource and low-resource languages to train the model. English and Japanese are the primary languages in the VGSAlign framework and have shown promising results. Our goal is to further train and improve these languages to create a diverse multilingual VSG model. Due to a lack of VGS models compatible with other languages, we use an English-based pre-trained model for training in low-resource languages, specifically Vietnamese and Indonesian. Despite the fact that English and Japanese are not considered low-resource languages, their inclusion is due to the availability of resources such as pre-trained models and the Flickr8K audio dataset. This allows for comparisons and benchmarking against these extensively studied languages.

The experimental results indicate that we can distinguish multilingual speech and image representations. The multilingual speech representations are distinct in the left area, while multilingual image representations are found in the right area. As for multimodal representation, the image and speech representations of English and Japanese are closely related, whereas those of Indonesian and Vietnamese are considerably distant.

The improved results on the VGS datasets, achieved by using Flickr8K, to find image-speech pairs without relying on text, suggest that our VGS Models for four languages have a promising approach in contributing to the field of multilingual self-supervised Visually Grounded Speech Models. These models also show potential in performing well on other languages that lack paired and transcribed data, thanks to their ability to learn speech representations from unlabeled data. Table 1 demonstrates the capability of the self-supervised VGS models to learn co-representation and effectively determine the similarity between speech and its corresponding image. This ability is crucial for aligning speech from multiple languages.

5. Conclusion

In conclusion, this research has successfully achieved promising results in multilingual self-supervised VGS models in four languages: EN, JA, ID, and VN. This was accomplished by employing fine-tuning and further training strategies on the based VGS models in VGSAlign. These models have been validated and evaluated using the Speech-Image Retrieval Recall Score, which demonstrates their ability to retrieve image-speech pairs without relying on text.

In the future, we plan to develop speech alignment for the four languages. The output of our multilingual VGS models will be used as input to compute the similarity between each speech, enabling us to determine pairs of related speeches for two source and target languages.

6. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681, as well as JST Sakura Science Program.

7. Bibliographical References

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. 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)*, pages 613–622, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. *Cross-lingual language model pretraining*. Curran Associates Inc.

Google. Text-to-Speech AI. https://cloud.google.com/text-to-speech?hl=en.

David Harwath, Galen Chuang, and James Glass. 2018. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973. IEEE.

David Harwath and James Glass. 2017. 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)*, pages 506–517, Vancouver, Canada. Association for Computational Linguistics.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. *Advances in Neural Information Processing Systems*, 29.

G Hinton and L van der Maaten. 2008. Visualizing data using t-sne journal of machine learning research.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. In *Proc. Interspeech 2019*, pages 1123–1127.

Herman Kamper and Michael Roth. 2018. Visually Grounded Cross-Lingual Keyword Spotting in Speech. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 253–257.

Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. 2017. Visually grounded learning of keyword prediction from untranscribed speech. In *Interspeech*.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2020. End-to-end speech translation with transcoding by multi-task learning for distant language pairs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1342–1355.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022. 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*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Xinjian Li, Ye Jia, and Chung-Cheng Chiu. 2023. Textless direct speech-to-speech translation with discrete speech representation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Satoshi Nakamura. 2009. Overcoming the language barrier with speech translation technology. Technical report, Citeseer.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The atr multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

Luan Thanh Nguyen and Sakriani Sakti. 2023. Vgsalign: Bilingual speech alignment of unpaired and untranscribed languages using self-supervised visually grounded speech models. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 53–57.

Junrui Ni, Liming Wang, Heting Gao, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark A. Hasegawa-Johnson. 2022. Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition. In *Interspeech*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Puyuan Peng and David Harwath. 2022. Word discovery in visually grounded, self-supervised speech models. In *Interspeech*.

Hyeonggon Ryu, Arda Senocak, In So Kweon, and Joon Son Chung. 2023. Hindi as a second language: Improving visually grounded speech with semantically similar samples. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Tohru Shimizu, Yutaka Ashikari, Eiichiro Sumita, Jinsong Zhang, and Satoshi Nakamura. 2008. Nict/atr chinese-japanese-english speech-to-speech translation system. *Tsinghua Science and Technology*, 13(4):540–544.

Changhan Wang, Hirofumi Inaguma, Peng-Jen Chen, Ilia Kulikov, Yun Tang, Wei-Ning Hsu, Michael Auli, and Juan Pino. 2023. Simple and effective unsupervised speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10771–10784, Toronto, Canada. Association for Computational Linguistics.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

8. Language Resource References

Pham Thanh Trung. 2022. Flickr8k Vietnamese Captions. https://www.kaggle.com/datasets/trungit/flickr8k-vi-caps.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.

Herman Kamper, Mark Hasegawa-Johnson. 2018. flickr. https://github.com/JSALT-Rosetta/flickr.

A. A. Nugraha, A. Arifianto, and Suyanto. 2019. Generating image description on indonesian language using convolutional neural network and gated recurrent unit. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6.