INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Weakly-supervised Speech-to-text Mapping with Visually Connected Non-parallel Speech-text Data using Cyclic Partially-aligned Transformer

Johanes Effendi[1], *Sakriani Sakti*[1,2], *Satoshi Nakamura*[1,2]

[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advance Intelligence Project (AIP), Japan
{johanes.effendi.ix4, ssakti, s-nakamura}@is.naist.jp

Abstract

Despite the successful development of automatic speech recognition (ASR) systems for several of the world's major languages, they require a tremendous amount of parallel speech-text data. Unfortunately, for many other languages, such resources are usually unavailable. This study addresses the speech-to-text mapping problem given only a collection of visually connected non-parallel speech-text data. We call this "mapping" since the system attempts to learn the semantic association between speech and text instead of recognizing the speech with the exact word-by-word transcription. Here, we propose utilizing our novel cyclic partially-aligned Transformer with two-fold mechanisms. First, we train a Transformer-based vector-quantized variational autoencoder (VQ-VAE) to produce a discrete speech representation in a self-supervised manner. Then, we use a Transformer-based sequence-to-sequence model inside a chain mechanism to map from unknown untranscribed speech utterances into a semantically equivalent text. Because this is not strictly recognizing speech, we focus on evaluating the semantic equivalence of the generated text hypothesis. Our evaluation shows that our proposed method is also effective for a multispeaker natural speech dataset and can also be applied for a cross-lingual application.

Index Terms: Speech-to-text mapping, non-parallel data, weakly-supervised, vector-quantized variational autoencoder, cyclic partially-aligned Transformer.

1. Introduction

Most speech-to-text transformation systems within a language are assumed to be automatic speech recognition (ASR), which transcribes the content spoken utterance into text sequences. The state-of-the-art ASR technologies with deep learning frameworks have been shown to reach human parity in performance [1, 2]. However, these systems are mostly trained using supervised learning paradigms that rely on a huge amount of parallel speech-text data. Although such a training technique can be applied for 10-20 of the world's major languages, this approach is not feasible for many other languages where gathering such a huge data collection is impossible [3].

Many researchers are aware of this problem, and several attempts on learning style have been made to reduce the quantity of parallel data required. One way is to train an ASR system in a semi-supervised manner using paired and unpaired speech-text data [4, 5, 6]. One semi-supervised approach utilized cycle consistency within the speech chain framework [7, 8] which enabled ASR and text-to-speech synthesis (TTS) to support each other given unpaired speech-text data. Another variant of cycle-consistency training used an alternative text-to-encoder model [9]. Effendi et al. (2020) proposed a multimodal machine chain that further reduced the need for paired and unpaired speech-text data by enabling a cross-modal augmentation from unrelated modality data. Specifically, the framework successfully improved ASR performance using additional image data.

Recently, Liu et al. (2020) proposed another alternative based on semi-supervised speech recognition that applied quantized-speech representation learning. Unfortunately, although they claimed that their work was a path toward unsupervised ASR, their proposed method still relied on paired speech-text data to train the initial model. On the other hand, Pasad et al. (2019) proposed a semantic text retrieval system through a multi-task learning mechanism that leveraged visual grounding. Similarly, their proposed framework still relied on paired speech-text data to build a shared representation. Consequently, all of these existing works still rely on a certain amount of paired data to initially train the model.

In human communication, on the other hand, it often does not matter whether we can figure out word-by-word what the speaker is saying as long as we understand the semantic message the speaker wants to convey. Therefore, we argue that it may be possible to address the construction of spoken language processing without having speech utterances and the exact corresponding transcriptions, which are generally unavailable. In fact, there are many available collections of texts and pictures from online books, and there are many available speeches recorded with images/videos in social media (i.e., YouTube). If we could link to those images, we might be able to create visually connected non-parallel speech-text data.

This study addresses weakly-supervised speech-to-text mapping problem given only a collection of visually connected non-parallel speech-text data. This may be considered one of the new ways of building speech-to-text transformation systems within a language but without using ASR. The system learns the semantic association between speech and text instead of recognizing the content of speech utterances with an exact word-by-word transcription. It can also be considered as a paraphrasing or translation task from unknown untranscribed speech utterances into semantically equivalent texts. Since this system does not strictly recognize speech, we focus on evaluating the semantic equivalence of the generated text hypothesis.

2. Related Works

Recently, research on constructing technologies with purely non-parallel data has been gained attention. To date, various approaches have been proposed for developing voice conversion systems with non-parallel data [10, 11, 12, 13]. One approach applies unsupervised neural machine translation to develop a text-to-text translation system without using any paired data [14, 15, 16, 17]. However, those works focus on mapping within a single modality framework (i.e., speech-to-speech or text-to-text). On the other hand, mapping between different modalities is more challenging due to the differences in the data

http://dx.doi.org/10.21437/Interspeech.2021-970

characteristics.

In a speech-to-text mapping task, speech features are continuous vector sequences while the corresponding text is formed in discrete sequences. Unfortunately, scant research has considered multi-modality mapping tasks with non-parallel data. Within the limited research on speech-to-text mapping tasks with non-parallel data, Sarl et al. (2020) recently proposed a spoken language understanding system trained on non-parallel speech and text data [18]. However, the model is more focused on dialog-act recognition rather than generating a descriptive sentence.

In this study, we focus on generating a descriptive sentence of the message being spoken. Specifically, the system attempts to learn how to generate semantically related text messages from speech utterances. We introduce the possibility of conducting weakly-supervised learning based on non-parallel data using a partially-aligned Transformer. We also introduce discrete speech representations using a vector-quantized variational autoencoder (VQ-VAE) to reduce the complexity of speech-text mapping, which also solves the low-resource problem and opens up possibilities for our proposed method to be used in an untranscribed unknown language.

3. Proposed Method

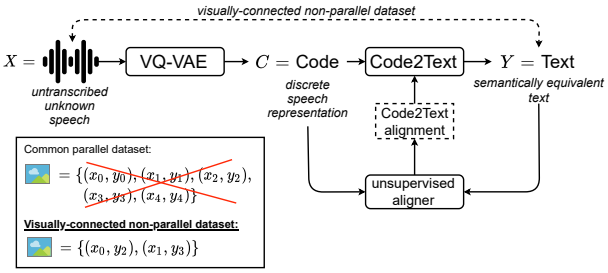

Figure 1: *Overview of proposed system.*

Our proposed framework transforms a speech X into a sentence Y, by leveraging the non-parallel speech and text data (Figure 1). First, to simplify the speech variability and its length discrepancies in text, we train a Transformer-based VQ-VAE to learn a discrete speech representation in a self-supervised manner. Then, we perform unsupervised alignments between the resulting discrete speech representation and the discrete target text sequences. Since the speech source and target text are generally based on the same images, we assume that some parts of speech and text content are semantically associated or aligned, which are then used by a partially-aligned Transformer model for speech-text mapping. Finally, we use the cycle mechanism as an augmentation to further improve the partially-aligned Transformer model.

3.1. Transformer-based Vector-quantized Variational Autoencoder

We use the VQ-VAE model [19] with transformer-based encoder and decoder [20] which was shown to provide good discretization performance for untranscribed unknown languages in the recent Zero Resource Speech Challenge [21, 22]. This model M_{vq} learns the discrete code representation c, so that $c = M_{vq}(x)$.

A VQ-VAE consist of an encoder and decoder, with a vector-quantizer module between them. Here, the training objective is defined as:

$$L_{VQ} = -\log p_\phi(x|z,s) + \|\text{sg}(z) - C\|_2^2 + \gamma\|z - \text{sg}(C)\|_2^2, \quad (1)$$

where function $\text{sg}(\cdot)$ stops the gradient, defined as:

$$x = sg(x); \quad \frac{\partial\, \text{sg}(x)}{\partial\, x} = 0. \quad (2)$$

L_{VQ} consists of three parts. The first part $-\log p_\phi(x|z,s)$ is a negative log-likelihood reconstruction loss between the input speech feature and the generated speech feature. The second part $\|\text{sg}(\hat{z}) - C\|_2^2$, is used to ensure that codebook C is close to the encoded representation z. Finally, the third term, $\|z - \text{sg}(C)\|_2^2$, updates the encoder.

3.2. Partially-aligned Code2Text Transformer Model

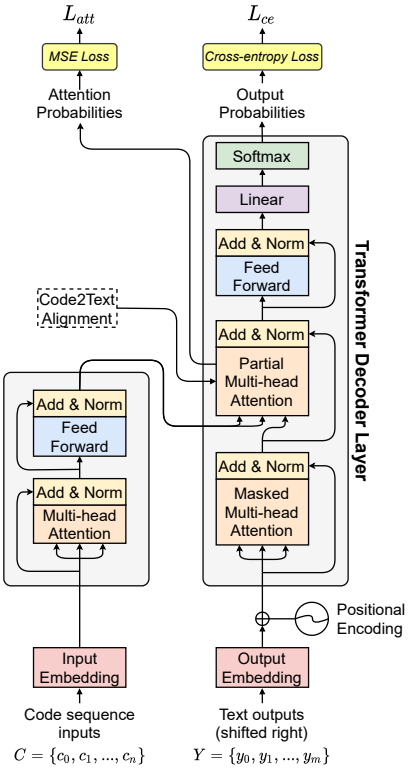

Figure 2: *A Transformer-based Code2Text for partially-aligned input-output.*

A partially-aligned Code2Text model uses the alignment of discrete speech representation $C = \{c_0, c_1, \ldots, c_n\}$ with the discrete target text sequences $Y = \{y_0, y_1, \ldots, y_m\}$. Inspired by the partially-aligned training strategy [23] for sequence-to-sequence neural machine translation (NMT), we modified a vanilla Transformer-based NMT model [24] into a partially-aligned Code2Text Transformer model by leveraging the alignment information between the input and output (see Fig. 2). Let us assume that P_c and P_y form the list of aligned words from the C and Y sequences. First, we penalized the source-to-target attention score in the decoder, so if $y_j \notin P_y$, the attention context vector for that word is zero ($C_t = 0$). Then, we also add an additional attention loss to emphasize the alignment between the partially-aligned part in a supervised manner. We create a hard-attention matrix H, where:

$$H_{i,j} = \begin{cases} 1 & \text{if } c_i \in P_c \text{ and } y_j \in P_y \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

so that the original attention matrix A can be supervised with attention loss L_{att} as follows:

$$L_{att} = \sum_{i=0}^{n}\sum_{j=0}^{m} \|A_{i,j}, H_{i,j}\|_2^2. \quad (4)$$

Finally, we weighted the softmax cross-entropy loss L_{ce} with L_{att} as follows:

$$L = L_{ce} + \alpha L_{att}. \quad (5)$$

3.3. Cycle Mechanism

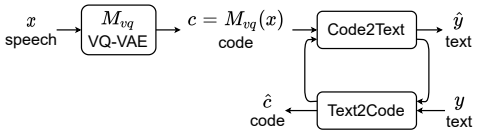

Figure 3: *Unsupervised augmentation with chain mechanism.*

Inspired by the success of the chain mechanism as an augmentation strategy in a cross-modal model [7, 25, 26], we implemented a Code2Text and Text2Code chain to further improve the performance of our proposed method (Figure 3). Given a text-only dataset D_y, a text y is translated using the Text2Code model, generating a $\hat{c}$ code hypothesis. This code hypothesis is then translated back into $\hat{y}$ by the Code2Text model. Then, we can backpropagate the Code2Text model using the reconstruction loss between y and $\hat{y}$.

4. Experiment Settings

4.1. Dataset

We used Flickr8k [27], which contains 8k images of everyday activities and events. For the synthetic speech caption, we generated single-speaker speech using GoogleTTS from the text caption. Then, for the natural speech caption, we used Flickr Audio [28], which was recorded using the crowdsourcing method with 183 unique speakers. We use the development and test sets which consist of 1k images each.

We formulated the training set differently from the original Flickr8k dataset by splitting the data as a visually grounded paraphrase (VGP) [29] to ensure a "semantic equivalence". In this task, we need to show how our proposed method can learn from non-parallel speech-text data but with a semantically similar meaning. While each image in this dataset has five speech and text captions, we choose two captions as speech only data, and another two captions as text-only data (Figure 1). Therefore, both the speech and text have the same image, which guarantees semantic equivalence between the pseudopair. This partition yields 12k speech utterances, and 12k of text caption, with both sets are disjointed ($(x_i, y_j), i \neq j$).

To show that our proposed approach can also be applied for a cross-lingual application, we also ran a cross-lingual experiment using English speech from SpeechCOCO multispeaker dataset [30] to the Japanese text from the STAIR Caption dataset [31], where this non-parallel speech-text mentions the same image from the MSCOCO dataset [32]. We take the matching amount of data and process it similarly to how Flickr8k dataset is handled. We assume that the English data is speech from an unknown and untranscribed language, where a visually connected non-parallel Japanese text exists.

4.2. Model Parameters

We extracted the Mel-spectrogram (80 dimensions, 25-ms window size, 10-ms time steps) using the Librosa package [33]. This speech feature is used as the input and output of the VQ-VAE model that has a 256 codebook size and 32 code dimensions. For adapting with the natural speech dataset, we froze the codebook part of the VQ-VAE model so that each code still represented the same speech segment.

We used a Transformer-based text encoder and decoder with a depth of 6 and a size of 512 hidden units. For the output layer, we used label smoothing with a factor of 0.005 and beam decoding with a size of 3. The vocabulary consists of words in the text-only training data that appear at a frequency of more than one time. We used Fast Align [34] as the unsupervised aligner. Inside the chain mechanism, we only updated the last element of the chain due to memory limitation. We trained all models with the Adam optimizer [35] using a learning rate of 1e-4.

4.3. Evaluation Method

We evaluate our proposed model performance by running an inference step using the dev and test set of the dataset the model is trained with. We use common metrics in the image captioning task: bilingual evaluation understudy (BLEU) with 4-gram [36] and CIDEr [37]. A BLEU score measures the n-gram similarity between the hypothesis and references, while CIDEr measures consensus by evaluating beyond the n-gram exact similarity. We used both metrics in a multi-reference condition. BLEU score is measured in percentage (i.e. multiplied by 100). In addition, to evaluate the semantic aspect, we developed a cosine-similarity based metric (Sim%) for multi-reference evaluation by calculating the highest cosine similarity between the hypothesis sentence embedding and the reference sentence embeddings. We generated the sentence embedding using the Sentence Transformers toolkit [38] with the pretrained models of RoBERTa [39] for English and Universal Sentence Encoder [40] for Japanese.

We calculated the corpus vocabulary statistics such as number of unique words and the vocabulary utilization ratio to measure how rich the hypotheses are. We reported the Pearson's correlation coefficient (r) score between the word frequencies of the hypothesis and the training set to show how good a model could learn to mimic the training set's word distribution.

5. Result and Discussion

Table 1: *Experiment result in the Flickr8k synthesized speech non-parallel dataset.*

Model	Dev			Test		
	Sim%	BLEU	CIDEr	Sim%	BLEU	CIDEr
(Baseline)						
Random selection	16.73	2.28	3.42	16.25	2.22	3.58
ASR [41]	16.86	5.64	7.63	16.94	4.69	7.08
(Proposed)						
Code2Text	35.58	15.30	31.48	35.79	15.04	31.66
+Partial Code2Text	40.58	16.95	36.11	40.94	16.80	36.86
+Cycle Augmentation	40.03	16.74	36.44	**40.47**	**17.25**	**37.52**

Table 2: *Adapting best Speech2Text model trained on Table 1 to the Flickr8k multispeaker natural speech non-parallel dataset.*

Model	Dev			Test		
	Sim%	BLEU	CIDEr	Sim%	BLEU	CIDEr
(Baseline)						
ASR [41]	16.30	3.23	9.30	15.18	3.09	9.07
(Proposed)						
Cyclic Partial Code2Text						
no adaptation	21.37	7.84	11.64	21.31	7.83	11.69
with adaptation	35.70	14.64	29.80	**35.35**	**14.57**	**29.01**

Table 3: *Our proposed Speech2Text vocabulary utilization statistics for the Flickr8k multispeaker natural speech dataset (Table 2) in comparison to the baseline.*

Metric	Baseline	Proposed
Number of unique words	20	300
Vocab utilization ratio	0.69%	10.42%
Pearson correlation (r)	0.343	0.958

In Table 1, we provide the baseline score of a random selection to show that our trained model produces a coherent hypothesis. We also reported the score of the ASR model trained directly on the non-parallel speech-text data. Then, we trained the VQ-VAE model using the speech data, and generated the code sequence as a discrete speech representation. The code sequence can then be used to train a Code2Text model against the partially-aligned text caption. Our proposed Code2Text model delivers a better score than the ASR baseline, which shows that our discretization method using VQ-VAE provides more efficient learning due to reduced variability compared with mel-spectrogram.

Then, because the input and output are discrete, we can approximate the alignment between the generated code sequence and the partially-aligned text using an unsupervised aligner. We next use the alignment information to influence the source-to-target multi-head attention by producing an additional L_{att}. We found that by multiplying L_{att} with $\alpha = 0.9$, we could obtain about 5.15% cosine similarity and 5.2 CIDEr points improvement on the test set, compared with a no-alignment model (Code2Text). We also trained the partially-aligned Text2Code with the same steps. After that, we use it in a cycle mechanism to achieve cross-modal augmentation which yielded a 0.66 CIDEr improvement. We also trained an ASR model with parallel data for a topline comparison, which yield a 89.81% cosine similarity, 81.43 BLEU, and 206.59 CIDEr scores on the test set.

Furthermore, we adapted our trained model to also support a multispeaker natural speech dataset using Flickr8k multispeaker natural speech dataset, in which we also use for testing. As shown in Table 2, our adaptation improves CIDEr by 20 points compared to the baseline ASR and 17 points compared to simply using the best model in Table 1 (no adaptation). We also trained a topline model with parallel dataset, which yield 82.75% cosine similarity, 70.24 BLEU, and 176.42 CIDEr scores. Next, we took the best score of the test set from Table 2 and compared the corpus statistics in Table 3. We found that the baseline system did not converge, as shown by the very low number of unique words with only 0.69% of the vocabulary being used. In comparison, our proposed model yielded 10.42% vocabulary utilization ratio. Moreover, our proposed method shows a better modelling of the vocabulary with a Pearson correlation (r) of 0.958, which is close to the topline of 0.999. This shows that our proposed partially-aligned Code2Text can model the training set word distribution as successfully as the topline, even without using any parallel data. In addition, even with limited vocabulary, our proposed method can still effectively convey the semantics of a partially-aligned speech.

Table 4: *Example results from the test set (Table 2).*

Model	Sentence
Baseline	two dogs are running through the grass .
Proposed	a woman and a little girl are smiling .
Reference	a laughing woman holding a little girl .
Baseline	a man and woman pose for a picture .
Proposed	a man in a red shirt is rock climbing .
Reference	a man poses as he jumps from rock to rock in a forest .

Table 4 shows a comparison of results between our proposed model, baseline ASR, and the input speech transcription (reference). The first example shows the baseline ASR model hypothesis which is totally unrelated to the reference. Our proposed method generated a hypothesis that semantically, closely resembles the reference, even while replacing the word "laughing" with "smiling". Then, in the second example, our proposed method successfully described the rock-climbing activity mentioned in the speech (reference). Although it is not an exact one-to-one transcription, the speech content itself can be successfully described in each of our proposed method's generated hypotheses. We are confident that this result will be very useful under the condition where no parallel speech-text data are available, in addition to handling an untranscribed unknown speech language.

Table 5: *Experiment result under cross-lingual EN-JA condition of transforming multispeaker English speech [30] to non-parallel Japanese text [31].*

Model	Dev			Test		
	Sim%	BLEU	CIDEr	Sim%	BLEU	CIDEr
ASR [41]	24.85	2.39	1.63	25.13	2.50	1.54
(Proposed)						
Code2Text	30.15	13.17	12.96	30.28	13.49	13.22
+Partial Code2Text	30.08	13.33	13.94	30.06	13.41	13.57
+Cycle Augmentation	30.51	13.36	14.21	30.33	13.40	13.75

Finally, we demonstrate how our proposed method can be used under a cross-lingual condition. As shown in Table 5, we found that the partial Code2Text and the cycle augmentation showed a little improvement in terms of CIDEr score. We hypothesize that this is due to the difficulty of aligning between different language structures (i.e., SVO for English, but SOV for Japanese). Nevertheless, while the baseline ASR did not show convergence, our proposed model could still achieve BLEU score of about 13 points even with a small amount of non-parallel data. This shows the effectiveness of our proposed discretization using the transformer-based VQ-VAE.

6. Conclusion

In this study, we investigated a weakly-supervised mapping task to transform unknown untranscribed speech utterances into a semantically equivalent text, even without a parallel speech-text dataset. Our proposed system uses a pipeline of VQ-VAE to generate a discrete speech representation, and a partially-aligned Code2Text Transformer model to learn the mapping between the code and the text. We also employed a cyclic augmentation strategy to further improve the performance of the Code2Text model. Our experiments with a multispeaker natural speech dataset showed improvement in every aspect that we examined. Our analysis of the text hypothesis shows that our proposed method can produce a more semantically relevant text. For future work, we will explore methods to increase the vocabulary utilization ratio, including an adversarial training method.

7. Acknowledgements

Part of this work is supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP17H06101 and JP21H03467, as well as Google AI Focused Research Awards Program.

8. References

[1] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," in *Proc. INTERSPEECH*, 2017, pp. 132–136.

[2] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *Proc. IEEE ICASSP*, 2018, pp. 5934–5938.

[3] G. Adda, S. Stker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. V. de Velde, F. Yvon, and S. Zerbian, "Breaking the unwritten language barrier: The BULB project," in *Proc. SLTU*, 2016, pp. 8–14.

[4] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, and J. Cernocky, "Semi-supervised sequence-to-sequence asr using unpaired speech and text," in *Proc. INTERSPEECH 2019*, pp. 3790–3794.

[5] N. Moritz, T. Hori, and J. Le Roux, "Semi-supervised speech recognition via graph-based temporal classification," *arXiv preprint arXiv:2010.15653*, 2020.

[6] A. Xiao, C. Fuegen, and A. Mohamed, "Contrastive semi-supervised learning for ASR," *arXiv preprint arXiv:2103.05149*, 2021.

[7] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *Proc. IEEE ASRU*, 2017, pp. 301–308.

[8] ——, "Machine speech chain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976–989, 2020.

[9] T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux, "Cycle-consistency training for end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2019, pp. 6271–6275.

[10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-vc2: Improved cycleGAN-based non-parallel voice conversion," in *Proc. IEEE ICASSP*, 2019, pp. 6820–6824.

[11] Y.-C. Wu, P. L. Tobing, T. Hayashi, K. Kobayashi, and T. Toda, "The NU non-parallel voice conversion system for the voice conversion challenge 2018." in *Proc. Odyssey The Speaker and Language Recognition Workshop*, 2018, pp. 211–218.

[12] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE SLT*, 2018, pp. 266–273.

[13] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," *arXiv preprint arXiv:1907.10185*, 2019.

[14] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," *arXiv preprint arXiv:1710.11041*, 2017.

[15] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," *arXiv preprint arXiv:1804.07755*, 2018.

[16] S. Sen, K. K. Gupta, A. Ekbal, and P. Bhattacharyya, "Multilingual unsupervised nmt using shared encoder and language-specific decoders," in *Proc. ACL*, 2019, pp. 3083–3089.

[17] H. Bai, M. Wang, H. Zhao, and L. Li, "Unsupervised neural machine translation with indirect supervision," *arXiv preprint arXiv:2004.03137*, 2020.

[18] L. Sar, S. Thomas, and M. Hasegawa-Johnson, "Training spoken language understanding systems with non-parallel speech and text," in *Proc. ICASSP*, 2020, pp. 8109–8113.

[19] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. NIPS*, 2017, pp. 6306–6315.

[20] A. Tjandra, S. Sakti, and S. Nakamura, "Transformer VQ-VAE for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge," in *Proc. INTERSPEECH*, 2020, pp. 4851–4855.

[21] E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, L. Besacier, S. Sakti, and E. Dupoux, "The zero resource speech challenge 2019: TTS without T," in *Proc. INTERSPEECH*, 2019, pp. 1088–1092.

[22] E. Dunbar, J. Karadayi, M. Bernard, X. Cao, R. Algayres, L. Ondel, L. Besacier, S. Sakti, and E. Dupoux, "The zero resource speech challenge 2020: Discovering discrete subword and word units," in *Proc. INTERSPEECH*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4831–4835.

[23] Y. Wang, Y. Zhao, J. Zhang, C. Zong, and Z. Xue, "Towards neural machine translation with partially aligned corpora," *Proc. IJCNLP*, 2017.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[25] J. Effendi, A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking and visualizing: Improving ASR through multimodal chain," in *Proc. IEEE ASRU*, 2019, pp. 471–478.

[26] ——, "Augmenting images for ASR and TTS through single-loop and dual-loop multimodal chain framework," in *Proc. INTERSPEECH*, 2020, pp. 4901–4905.

[27] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL HLT*, 2010, pp. 139–147.

[28] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. IEEE ASRU*, 2015, pp. 237–244.

[29] J. Effendi, S. Sakti, K. Sudoh, and S. Nakamura, "Leveraging neural caption translation with visually grounded paraphrase augmentation," *IEICE Transactions on Information and Systems*, vol. 103, no. 3, pp. 674–683, 2020.

[30] W. Havard, L. Besacier, and O. Rosec, "SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set," in *Proc. GLU*, 2017, pp. 42–46.

[31] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale japanese image caption dataset," in *Proc. ACL*, Vancouver, Canada, 2017, pp. 417–421.

[32] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[33] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," 2015, pp. 18–24.

[34] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proc. HLT-NAACL*, 2013.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[37] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE CVPR*, 2015, pp. 4566–4575.

[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. EMNLP*, 2019.

[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[40] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Ábrego, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, "Multilingual universal sentence encoder for semantic retrieval," *arXiv preprint arXiv:1907.04307*, 2019.

[41] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 4960–4964.