From Speech Chain to Multimodal Chain: Leveraging Cross-modal Data Augmentation for Semi-supervised Learning

Johanes Effendi[1,2] Andros Tjandra[1] Sakriani Sakti[1,2] Satoshi Nakamura[1,2]
[1]Nara Institute of Science and Technology [2]RIKEN AIP
{johanes.effendi.ix4, andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

1 Introduction

Recently, approaches that utilize learning from feedback links which enable model training with unpaired datasets have gained attention. He et al. [1] and Cheng et al. [2] recently published work that proposed a mechanism called dual learning in neural machine translation (NMT). It can leverage monolingual data to improve NMT. Other similar systems in image modality such as CycleGAN [3] was also proposed. However, most only work with the same domain between the source and the target.

In speech processing, the speech chain framework [4, 5] was proposed to integrate human speech perception and production behaviors that utilize both automatic speech recognition (ASR) model, and text-to-speech synthesis (TTS) inside a chain mechanism. This approach enables semi-supervised training for both ASR and TTS, even without paired speech-text data, which is often unavailable. Perhaps this is the first framework that was constructed on a different domain (speech vs. text). However, this study is limited to speech and textual modalities.

In this study, we proposed a multimodal machine chain, which is an improvement from speech chain. We proposed a collaborative image captioning-retrieval model collaboration in visual chain, which works closely with the speech chain through text modality. Our new framework mimics the mechanism of the entire human communication system with auditory and visual sensors. The overall system and its comparison with speech chain can be seen in Fig. 1.

2 Training Mechanism

The sequence-to-sequence model in closed loop architecture allows us to train our entire model in a semi-supervised fashion by concatenating both the labeled and unlabeled data. To further clarify the learning process, we describe the mechanism based on the availability condition of the training data:

1. **Paired speech-text-image data exist:**
   Given complete multimodal dataset $D1$, all the models in the chain are trained separately supervisedly.
2. **Unpaired speech, text, images data exist:**
   In this case, speech, text, and image data are available in $D2$, but they are unpaired. The reconstructed data are used to calculate reconstruction loss for each respective model. For example, with the text-only dataset $D2_y$, TTS generates speech utterance $\hat{x}$ for the ASR. On the other hand, image captions y retrieve images $\hat{z}$, which are reconstructed into text $\hat{y}$ using the IC model.
3. **Single modality data exist:**
   In this case, only a single data (speech or image) is available, and the others are empty. For example, if only image data are available in dataset $D3_z$, first we perform unrolled process IC→IR in the visual chain (See 2(b)). The generated image caption $\hat{y}$ is then used to perform unrolled process TTS→ASR in the speech chain (See 2(c)).

We are trying to learn whether in the situation where only image data exist (3) we can still improve the ASR performance through a learning pro-

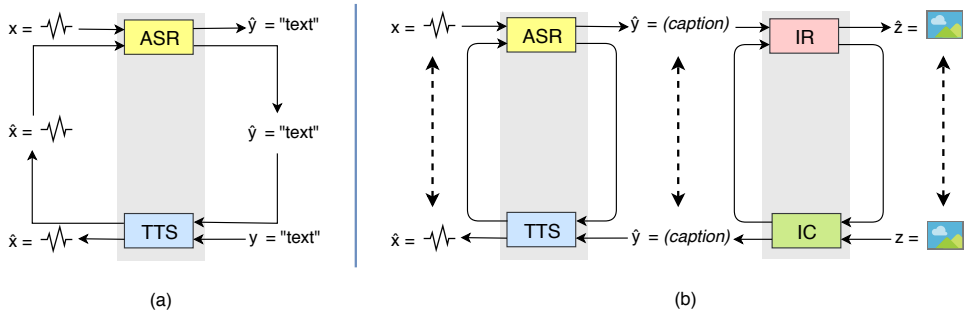

Figure 1: Architecture of (a) speech chain framework [4], and (b) our proposed multimodal chain mechanism.

cess from a visual chain to a speech chain by leveraging cross-modal data augmentation.

3 Experiment Set-up and Result

Table 1: Training data partition for Flickr30k with three conditions: (1) available paired data are denoted as ◯, (2) available data but unpaired are denoted as ▲, and unavailable data are denoted as ×.

Set	Speech x	Text y	Image z	#	Training Type
$D1$	◯	◯	◯	2000	1 (paired)
$D2$	▲	▲	▲	7000	2 (unpaired)
$D3_x$	▲	×	×	10000	3 (unpaired)
$D3_z$	×	×	▲	10000	3 (unpaired)

We ran our experiment with the Flickr30k dataset [6] that has 31,014 photos of everyday activities and scenes. Similar to other image captioning datasets, each image has five captions with a vocabulary of 18k words. To maintain balance between source and target, we selectively used one or five images depending on the task target. We generated speech from the Flickr30k captions using single speaker Google TTS. We used sequence-to-sequence encoder-decoder model for ASR [7], Tacotron [8], show-attend-tell model for IC [9], and Neural IR Embedding [10] for IR.

Table 2: ASR and TTS performance

Data	ASR WER(%)	TTS L2-norm^2
Baseline: ASR & TTS (Supervised learning - Type 1)		
$D1$ 2k*	81.31	0.874
Proposed: speech chain ASR→TTS and TTS→ASR (Semi-supervised learning - Type 2)		
$+D2$ 7k	10.60	0.714
Proposed: visual chain → speech chain (Semi-supervised learning - Type 3)		
$+D3z$ 10k	7.97	0.645
Topline: ASR & TTS separately (Supervised learning - Full Data)		
D_{all} 29k	2.37	0.398

Table 3: IC and IR performance

Data	IC BLEU1	IR R@10↑	IR med r↓
Baseline: IC & IR (Supervised learning - Type 1)			
$D1$ 2k*	33.91	26.88	34
Proposed: visual chain IC→IR and IR→IC (Semi-supervised learning - Type 2)			
$+D2$ 7k	42.11	28.14	31
Proposed: speech chain → visual chain (Semi-supervised learning - Type 3)			
$+D3_x$ 10k	43.08	28.44	30
Topline: IC & IR separately (Supervised learning - Full data)			
D_{all} 29k	66.27	62.42	5

Then, we demonstrate how much we can improve performance when the required data are no longer available, using our proposed chain mechanism. We used the initial model to train the speech chain using $D2$ 7k data and achieved 10.60% WER and 0.714 L2-norm^2 as shown in Table 2. Finally, using the IC model that was trained semi-supervisedly through Type 2(a)&2(c), we decoded the image-only $D3_z$ dataset which enables it to be used in speech chain. By this way, we achieved about 2.6% WER improvement over the original speech chain [4] that was only trained using the speech and text datasets. Our proposed strategies makes improvement of ASR and TTS possible, even without any speech or text data, with the help of a visual chain.

In the third block of Table 3 we show that the visual chain can also be improved using speech data, by the help of speech chain. There was about 1 point improvement in terms of BLEU for IC (high is good) and median r for IR (low is good). This result also implies that using our proposed learning strategy, the IC and IR model can be improved even without image and text datasets available. Therefore, we showed that it also works not only from image-to-speech modality, but also reversely.

4 Conclusions

We described a novel approach for cross-modal data augmentation that upgrades a speech chain into a multimodal chain. We improved the speech chain using an image-only dataset, bridged by our visual chain, and vice-versa. Therefore, we conclude that it is still possible to improve ASR, even without speech and text data available, with our proposed multimodal chain. In the future, we will jointly train both the speech and visual chain so that both can also be updated together.

5 Acknowledgement

Part of this work is supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237 as well as NII CRIS Contract Research 2019 and Google AI Focused Research Awards Program.

References

[1] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016.

[2] Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[4] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking: Speech chain by deep learning. *CoRR*, abs/1707.04879, 2017.

[5] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Machine speech chain with one-shot speaker adaptation. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, pages 887–891, 2018.

[6] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, Dec 2015.

[7] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

[8] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. 2017.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[10] Armand Vilalta, Dario Garcia-Gasulla, Ferran Parés, Eduard Ayguadé, Jesus Labarta, E Ulises Moya-Sánchez, and Ulises Cortés. Studying the impact of the full-network embedding on multimodal pipelines. *Semantic Web*, (Preprint):1–15.