Improving ASR with Multimodal Machine Chain*

◎ Johanes Effendi[1,2], Andros Tjandra[1,†], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]
([1]NAIST, [2]RIKEN AIP)

1 Introduction

Gathering a substantial amount of parallel data is one of the major problems in building a multimodal model. Previously, a machine speech chain [3] was proposed to enable the training of ASR and TTS to assist each other in semi-supervised learning. It successfully avoids the need for a large amount of paired speech and text data by using a speech or text only unpaired data in a closed-loop mechanism. However, this work has not yet addressed visual modality, although visual modality is also one of the important senses in human communication.

In this study, we formulate a multimodal model collaboration between automatic speech recognition (ASR), text-to-speech synthesis (TTS), image captioning (IC), and image generation (IG), in which they support each other under a weakly-supervised chain training strategy. The results reveal that with the help of IC and IG, ASR and TTS can still be improved using an image-only dataset.

2 Multimodal Machine Chain

We proposed a multimodal machine chain to mimic overall human communication and accommodate visual modality. This framework emphasizes that human communication is not only auditory but also visual. In addition to the previously published speech chain, we added a visual chain, which is composed as a collaboration between an image captioning model and an image generation model. We call this MMC1 for the multimodal chain with dual-loop architecture that uses text modality as a bridge (See Fig. 1).

Then, inspired by how human brain process multiple sensories in a unified manner, we also see the possibility to introduce sharing between ASR and IC, given that both of them has the same target modality, which is text. As an alternative, we also proposed a single-loop multimodal chain, which we call MMC2. With this alternative chain, we want to investigate the possibility of applying the chain mechanism in a simple multisource multimodal model.

3 Chain Components

For ASR and TTS models, we use a similar model as Tjandra et al. [3], which is based on a sequence-to-sequence architecture. Then, for IC, we removed the last two layers of ResNet model, and use the hidden representation to be attended by an attentional LSTM-based text decoder. We use the AttnGAN model for IG, which generates the image iteratively until the 128x128 pixels resolution. For ImgSp2Txt, we averaged the output layer of ASR and IC, to enables the combination of these two models in an ensembling fashion.

4 Experiment Settings

We used Flickr8k dataset [2] with Flickr8k Audio [1] which has 8000 photos of everyday activities and events. Each image has five captions with multi-speaker speech. We make a data partition from the 6000 images of the training set, into several subsets with different modality type. To simulate a small number of paired multimodal data, the first partition consists of 800 images, which is the smallest partition. The second partition is the unpaired multimodal data which has 1500 images. Finally, the last two partitions are the speech-only and image-only subset with 1850 images each. We trained the model step-by-step following the order of the partition (represented by every row in Table 1).

We evaluated each model with the commonly used automatic metric for each task. We used character or word error rate (CER/WER) to evaluate ASR and bilingual evaluation understudy (BLEU) with 4-grams (B4) for the IC. For TTS, we used L2-norm^2 metric to calculate the error between the generated mel-spectrogram and its reference. Finally, we used the inception score (IS) to measure how realistic the IG output was. Each model is evaluated with the

---

*Multimodal Machine Chain による音声認識の改善、
◎ジョハネスエフェンディ[1,2]、アンドロスチャンドラ[1,†]、サクリアニサクティ[1,2]、中村 哲[1,2]
([1]NAIST, [2]RIKEN AIP)
†This work was done when he was at NAIST, he is currently at Facebook AI, USA.

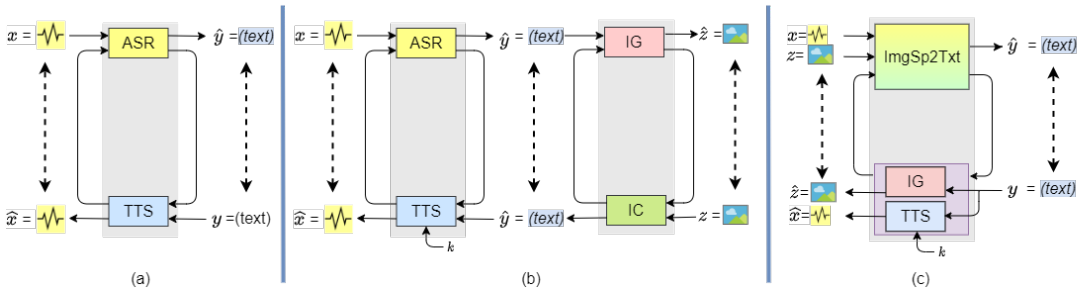

Fig. 1 Structure of: (a) speech chain [3] (b) proposed single-loop multimodal chain (MMC1), (c) proposed dual-loop multimodal chain (MMC2).

Table 1 Performance of proposed MMC1 and MMC2 compared with label propagation method in Flickr8k multispeaker natural speech dataset. The last line is the topline system when the 6k images with the corresponding five captions and five speech utterances for each image are available.

Training	Partition Type	#Image	MMC1				MMC2			
			ASR	**IC**	**TTS**	**IG**	**ImgSp2Txt**		**TTS**	**IG**
			CER↓	B4↑	$L2^2$↓	IS↑	CER↓	B4↑	$L2^2$↓	IS↑
Label Propagation (Semi-Supervised)	Multimodal (P)	800	36.35	12.75	0.77	5.90	26.67	32.23	0.77	5.90
	+ Multimodal (U)	1500	39.57	12.53	0.77	7.04	27.45	33.59	0.77	7.04
	+ Sp only (U)	1850	46.04	-	0.63	-	28.87	35.75	0.63	-
	+ Img only (U)	1850	-	11.41	-	7.20	30.31	35.38	-	7.20
Proposed Multimodal Chain (Semi-Supervised)	Multimodal (P)	800	36.35	12.75	0.77	5.90	26.67	32.23	0.77	5.90
	+ Multimodal (U)	1500	15.10	13.22	0.59	8.29	14.88	55.15	0.65	10.12
	+ Sp only (U)	1850	12.37	13.28	0.56	9.12	13.81	58.03	0.62	10.65
	+ Img only (U)	1850	12.06	13.29	0.56	9.11	12.32	59.66	0.61	9.95
Topline (Supervised)	Multimodal (P)	6000	5.76	19.91	0.50	9.66	5.16	79.88	0.50	9.66

test set of Flickr8k dataset.

5 Result and Conclusion

We used label propagation to compare our experiment results. Label propagation is a simple semi-supervised method that continues the training of a model with generated pseudo-labels from no-labelled data. In ASR case, for example, we can use an ASR model to transcript a speech only data, so that the transcription hypothesis can be used as a pseudo-label to continue the ASR training semi-supervised. However, the result in Table 1 shows that this method was not able to improve the ASR and ImgSp2Txt. We suspect that this is because this method needs more data to initialize the model training in a supervised manner.

The next part of Table 1 shows the result of our proposed method. We can see that with multimodal chain, the CER can be improved from 15.10% to 12.06% by using the speech only and image only data. The improvement can also be seen in ImgSp2Txt CER and B4 score, which performs better than MMC1 in low-data condition. We also observed that the performance of other tasks such as IC, TTS, and IG can be maintained. In conclusion, these results show that the cross-modal augmentation enabled by our proposed multimodal chain has shown to be effective to further improve the ASR model, even with an image only dataset.

6 Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Number JP17H06101 and the Google AI Focused Research Awards Program.

References

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

[2] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147, 2010.

[3] A. Tjandra, S. Sakti, and S. Nakamura. Listening while speaking: Speech chain by deep learning. In *Proc. of the IEEE ASRU*, pp. 301–308, Dec 2017.