Real-time Neural Machine Speech Chain *

◎ Sashi Novitasari[1], Andros Tjandra[1,†], Tomoya Yanagita[1],
Sakriani Sakti[1,2], Satoshi Nakamura[1,2]
([1]NAIST, [2]RIKEN AIP)

1 Introduction

Automatic speech recognition (ASR) and text-to-speech synthesis (TTS) systems are closely related to each other. Despite it, the development of these systems progressed less dependently. A machine speech chain framework was proposed for the semi-supervised training of the end-to-end attention-based ASR and TTS using a feedback loop [4], which was inspired by the human speech chain mechanism [1] for speech production. However, unlike the human speech chain, the current machine speech chain requires a long feedback and output delay due to the global attention mechanism.

In this work, we propose an incremental machine speech chain framework with a short-term feedback loop to reduce the delay in the machine speech chain. The challenge is to generate the feedback and output based on an incomplete input sequence with a small delay.

2 Machine Speech Chain

2.1 Basic Machine Speech Chain

Machine speech chain is a framework to train the attention-based sequence-to-sequence (seq2seq) ASR and TTS semi-supervisory by connecting them via a closed-loop [4]. It consists of two training phases: independent supervised training and joint unsupervised training. The independent supervised training phase acts as the knowledge initialization for ASR and TTS. After the ASR and TTS independent training, these systems are trained jointly through a closed-loop by using the unpaired speech and text data.

The ASR-TTS closed-loop consists of two unrolled processes: ASR-to-TTS and TTS-to-ASR. In ASR-to-TTS process, the ASR generates the transcription of an unlabeled speech utterance and the TTS reconstructs the speech using ASR text. In this process, we update the TTS model parameters based on the loss between the original speech and the TTS speech. A similar mechanism is also applied in the TTS-to-ASR process, given an unlabeled text sentence, to update the ASR model. By repeating these processes in the loop, ASR and TTS could improve together.

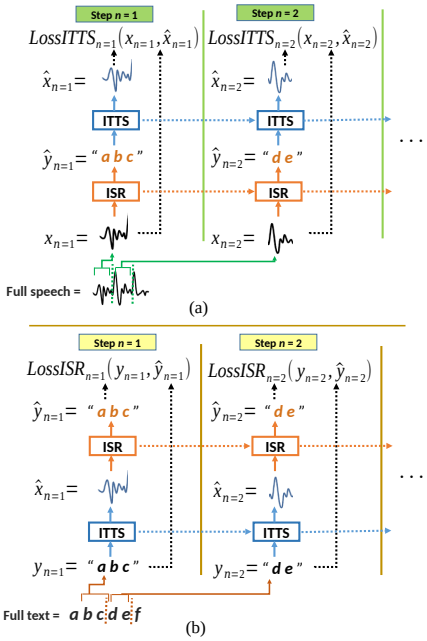

Fig. 1 Unrolled-loop processes in incremental machine speech chain: (a) ISR-to-ITTS and (b) ITTS-to-ISR.

2.2 Incremental Machine Speech Chain

The incremental machine speech chain follows the idea of the human speech chain and the basic machine speech chain for the joint construction of an incremental ASR (ISR) and an incremental TTS (ITTS). ISR and ITTS are systems that are capable of low-delay prediction. In the proposed framework, the ISR and ITTS are connected with a short-term closed-loop. Here the data passing between ISR and ITTS is done with a low delay without waiting for the complete input sequence.

The incremental machine speech chain is also consists of two training phases: ISR and ITTS inde-

---

* リアルタイムニューラルマシーンスピーチチェーン、
◎サシ ノビタサリ [1]、アンドロス チャンドラ [1,†]、柳田 智也 [1]、サクリア二 サクティ [1,2]、中村 哲 [1,2]
([1]NAIST, [2]RIKEN AIP)

† The work was done when he was at NAIST, he is currently at Facebook AI, USA.

Table 1 ASR and TTS performances on WSJ. (*nat-sp* = natural speech input; *nat-txt* = natural text input; *syn-sp* = TTS output as input; *rec-txt* = ASR output as input; *indep-trn*= independent training; *chain-trn-greedy* = joint training with greedy intermediate output generation; *chain-trn-teachforce*: joint training with teacher-forcing intermediate output generation).

Data	ASR (CER(%))				TTS (L2-norm^2)			
	Non-incremental (delay: 7.88 sec)		Incremental (delay: 0.84 sec)		Non-incremental (delay: 103 characters)		Incremental (delay: 30 characters)	
	nat-sp	**syn-sp**	**nat-sp**	**syn-sp**	**nat-txt**	**rec-txt**	**nat-txt**	**rec-txt**
ASR and TTS with independent training								
indep-trn (*SI-84*)	17.33	27.03	17.81	44.54	0.99	1.02	1.04	3.62
indep-trn (*SI-284*)	7.16	9.60	7.97	19.99	0.75	0.77	0.84	1.31
ASR and TTS with machine speech chain								
indep-trn (*SI-84*) + chain-trn-greedy (*SI-200*)	11.21	11.52	14.23	32.43	0.80	0.82	0.86	1.35
indep-trn (*SI-84*) + chain-trn-teachforce (*SI-200*)	7.27	6.30	9.43	12.78	0.77	0.80	0.79	1.26

pendent training and ISR-ITTS joint training with a closed-loop. In the independent training, we applied attention-transfer for ISR [2] and TTS, which models learned the incremental steps from the attention alignment of a standard seq2seq ASR for less-complex models construction. In the ISR-ITTS joint training, shown in Fig. 1, the output and feedback generations between the components are done based on the segment-level input. In each unrolled-loop process, the first component performs a prediction from a short part of the input, and the second component reconstructs the input segment that the first component receives. These processes are done incrementally until it reaches the last input segment.

3 Experiment

We performed the experiments on *Wall Street Journal* (WSJ) data [3] in Table 1. We used *SI-84* and *SI-284* sets for the systems independent training and *SI-200* set for the joint training. Our baselines are the ISR and ITTS that were trained independently using *SI-84*, while the toplines are the standard ASR and TTS that were trained with *SI-284* independently. We experimented on the natural and the synthetic data as the system input during inference. The synthetic data was the output of the first system in an unrolled-loop process. We also experimented on two approaches for intermediate output generation in each unrolled-loop process during training: greedy (unsupervised) and teacher-forcing (supervised).

Our incremental machine speech chain framework improved the ISR and ITTS performances with a shorter delay than the standard system. The ISR and ITTS improvements occurred on both with natural input and synthetic input. It implies that the short-term feedback loop between these systems leveraged their training quality.

4 Conclusion

We proposed an incremental macine speech chain that connects an ISR and an ITTS with a short-term closed-loop for the joint construction of these systems. It reduced the standard systems' delay and improved the ISR and ITTS performances with a close performance to the basic framework.

Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Number JP17H06101.

References

[1] P. B. Denes and E. N. Pinson. *The Speech Chain: The Physics and Biology of Spoken Language*. Science/communication. W.H. Freeman, New York, N.Y, 1993.

[2] S. Novitasari, A. Tjandra, S. Sakti, and S. Nakamura. Sequence-to-sequence learning via attention transfer for incremental speech recognition. In *Proc. Interspeech 2019*, pp. 3835–3839, 2019.

[3] D. B. Paul and J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, pp. 357–362. Association for Computational Linguistics, 1992.

[4] A. Tjandra, S. Sakti, and S. Nakamura. Listening while speaking: Speech chain by deep learning. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 301–308, 2017.