



Augmenting Images for ASR and TTS through Single-loop and Dual-loop Multimodal Chain Framework

Johanes Effendi^{1,2}, Andros Tjandra¹, Sakriani Sakti^{1,2}, Satoshi Nakamura^{1,2}

¹Nara Institute of Science and Technology, Japan

²RIKEN, Center for Advanced Intelligence Project (AIP), Japan

{johanes.effendi.ix4, andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

INTERSPEECH 2020

Outline

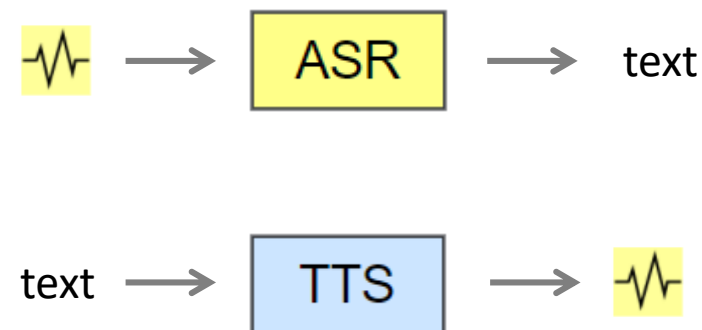


- Introduction
- Multimodal machine chain
- Experiment set-up
- Result and discussion
- Conclusion and future works

Introduction



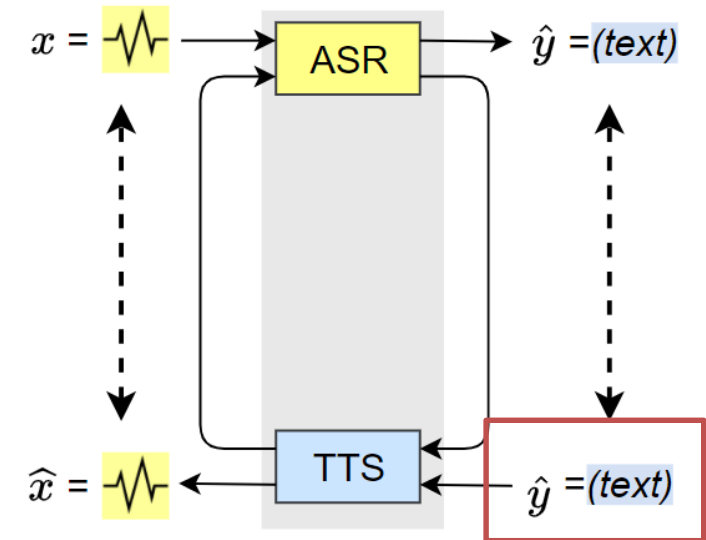
- Current state-of-the-art speech processing technology (i.e. ASR and TTS)
 - Rely on the availability of paired speech and transcription
- To improve: collect more data
- For some language, resources in such quantity are usually unavailable
- Some approaches to reduce data usage is needed



Speech Chain (Tjandra et al., 2017)



- Enables the training of ASR and TTS to assist each other in semi-supervised learning
- Avoids the need of large amount of paired speech and text data
- But still need a large amount of unpaired speech and text data
- Speech and text is the source and target modality of ASR and TTS



Can we improve ASR and TTS without speech or text data?

How human perceive senses



- Human communication channels is not only auditory but also visual
- Multiple information sources are perceived together
- Able to learn even when no paired data are available (less supervision)

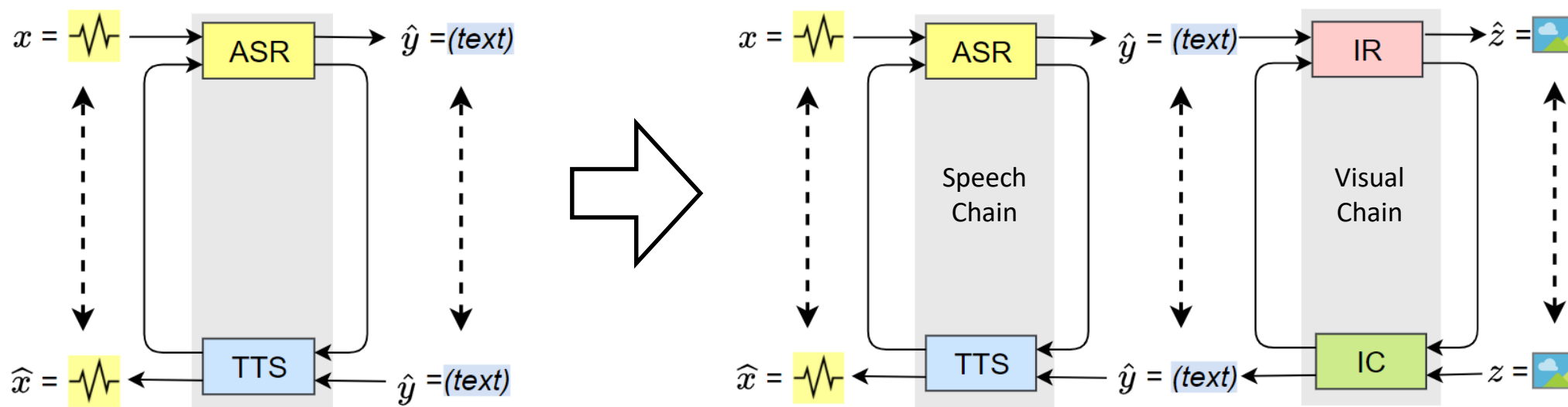


Learning by textual+visual



Learning by auditory+visual

Multimodal Machine Chain (Effendi et al., 2019)



- Proposed to mimic overall human communication and accommodate visual modality
- Speech chain (ASR+TTS) and visual chain (IC+IR)
- Evaluated on single-speaker synthesized speech
- IR model: difficulty to handle unseen images

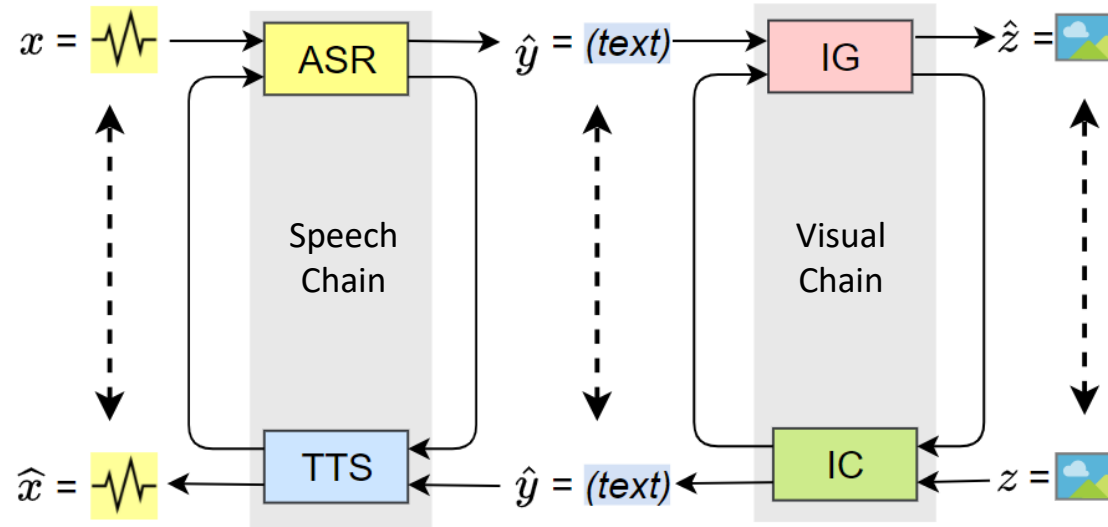
Our proposed model



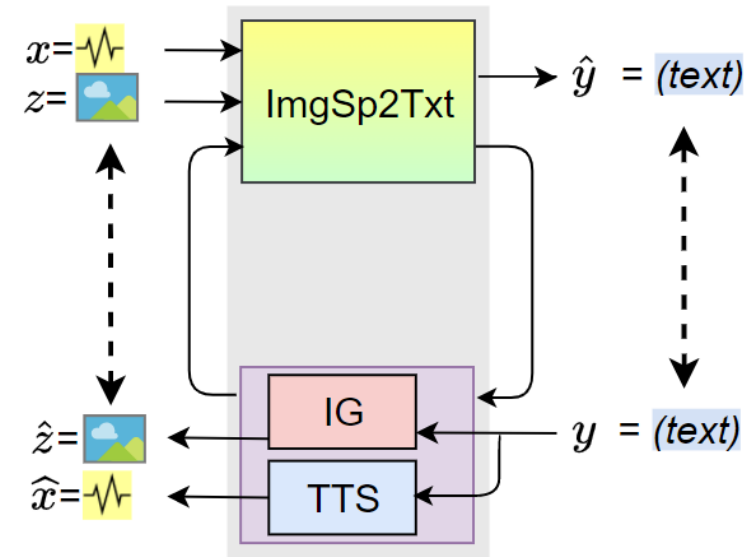
Components	Effendi et al. (2019)	This study
TTS	Single-speaker	Multi-speaker with one-shot speaker adaptation (Tjandra et al., 2018)
Evaluated on	Synthesized speech by Google TTS	Natural speech (Flickr8k Audio)
Image production	Image retrieval	Adversarial-based image generation
#loop	Dual-loop only	Single and dual-loop

- Image generation (IG) to handle unseen images
- Tested on multi-speaker natural speech dataset
- Multi-speaker TTS with embedding from DeepSpeaker (Li et al., 2017)
- One-shot speaker adaptation (Tjandra et al., 2018)

[Proposed] MMC1 and MMC2



Dual-loop MMC1



Single-loop MMC2

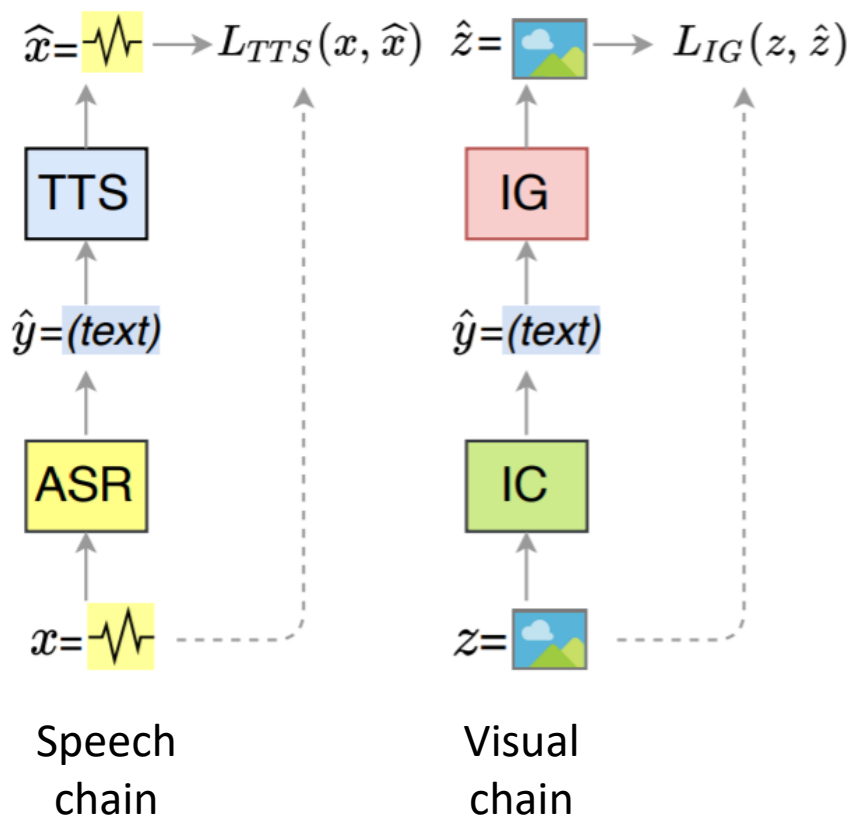
- MMC1: dual-loop architecture with text as the bridge
- MMC2: alternative for application example on multi-source multimodal model
- Human brain process visual and auditory components of speech in a unified manner (Calvert, 2001)
 - Introduce sharing between ASR and IC -> ImgSp2Txt

MMC1 unrolled process:

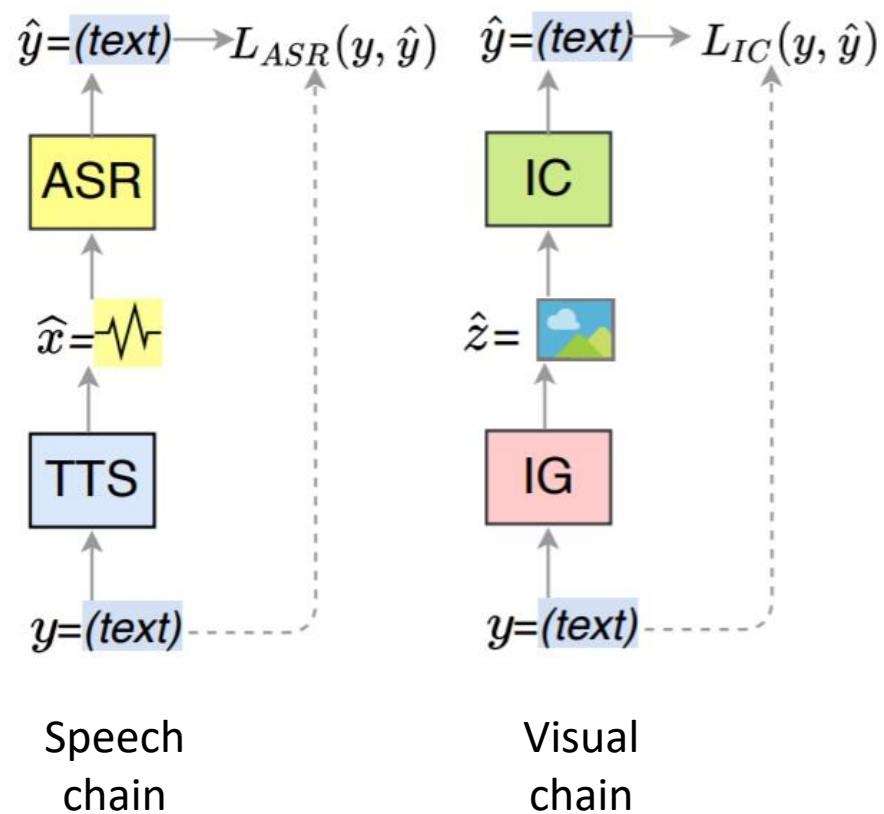
The loop inside the speech chain and visual chain



When the input is image or speech only data



When the input is text only data



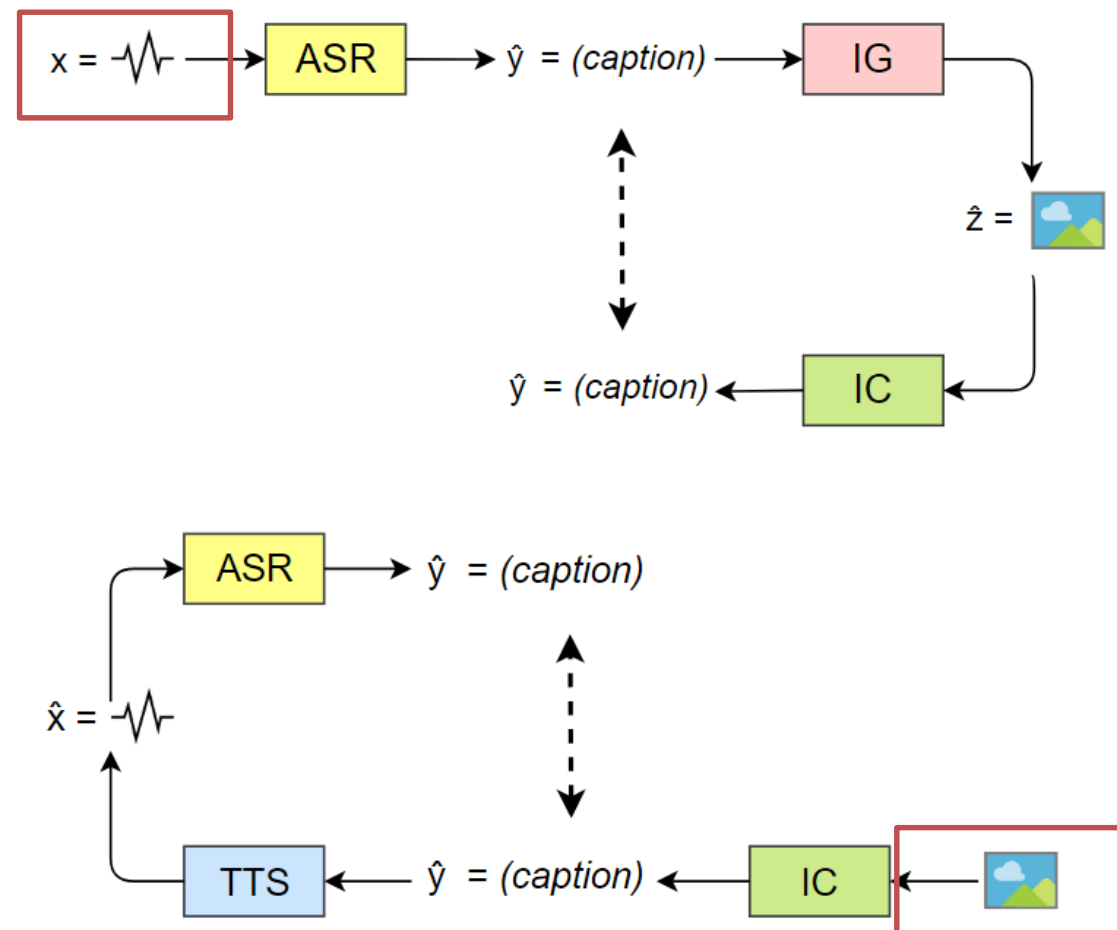
MMC1 unrolled process:

Speech chain and visual chain collaboration

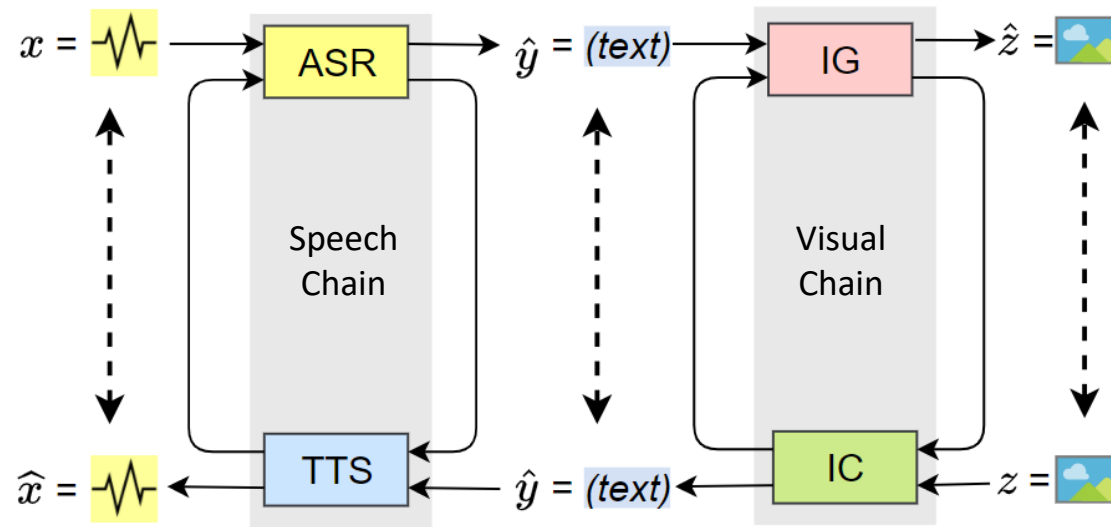


- Speech chain and visual chain collaborates through text modality
- The loss calculated from intermediate text
- Backpropagate the last element of the chain
- Simple filtering in text hypothesis

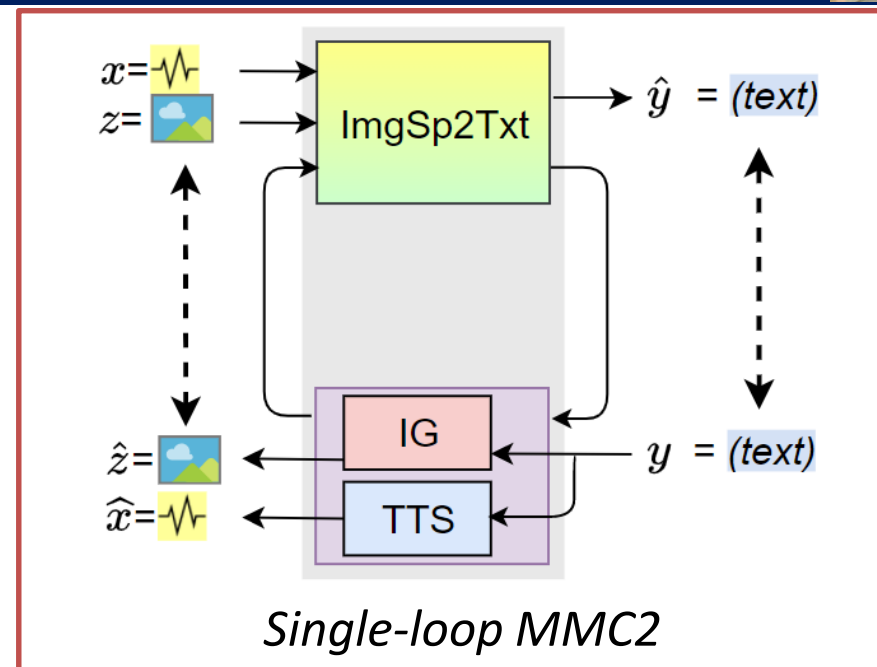
This is our main interest, to see if the image-only data can help improve ASR



[Proposed] MMC1 and MMC2



Dual-loop MMC1



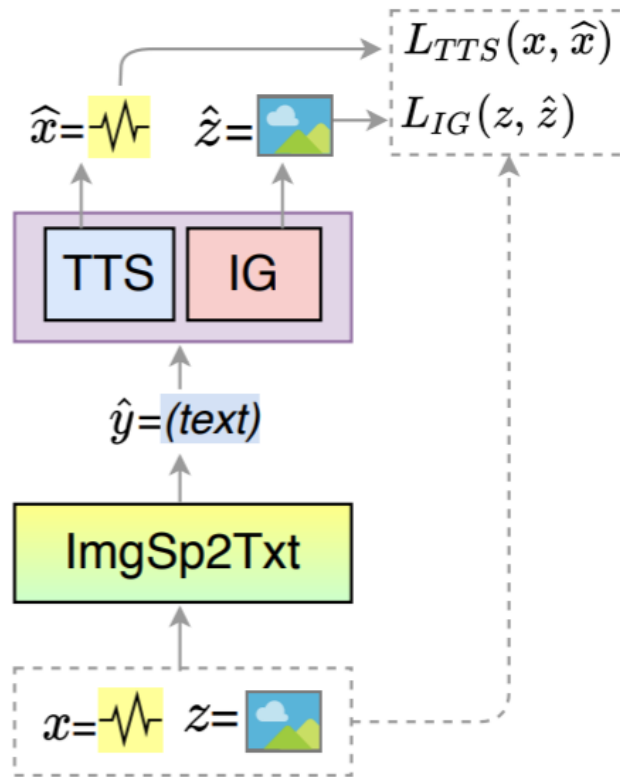
Single-loop MMC2

- MMC1: dual-loop architecture with text as the bridge
- MMC2: alternative for application example on multi-source multimodal model
- Human brain process visual and auditory components of speech in a unified manner (Calvert, 2001)
 - Introduce sharing between ASR and IC -> ImgSp2Txt

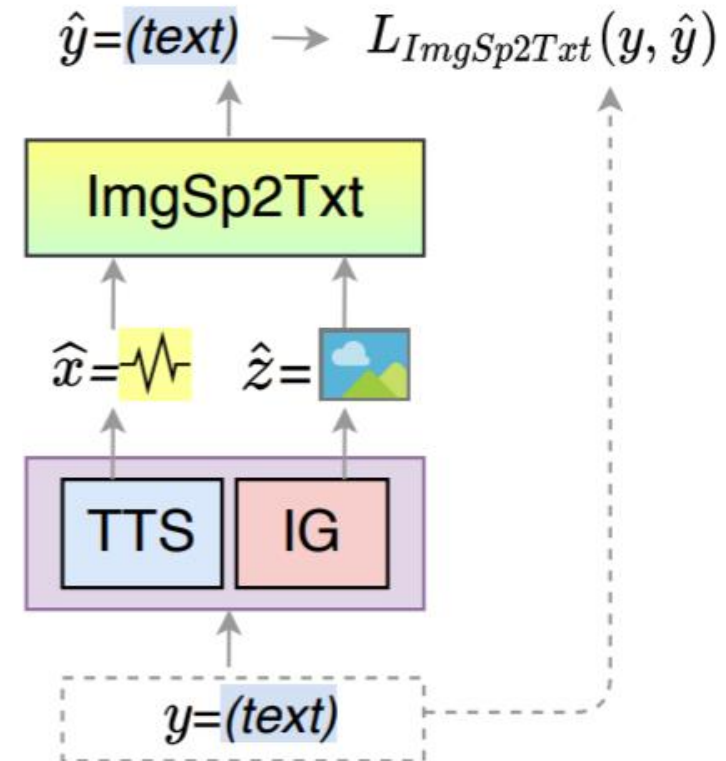
MMC2 unrolled process



When the input is image and/or speech only data



When the input is text only data



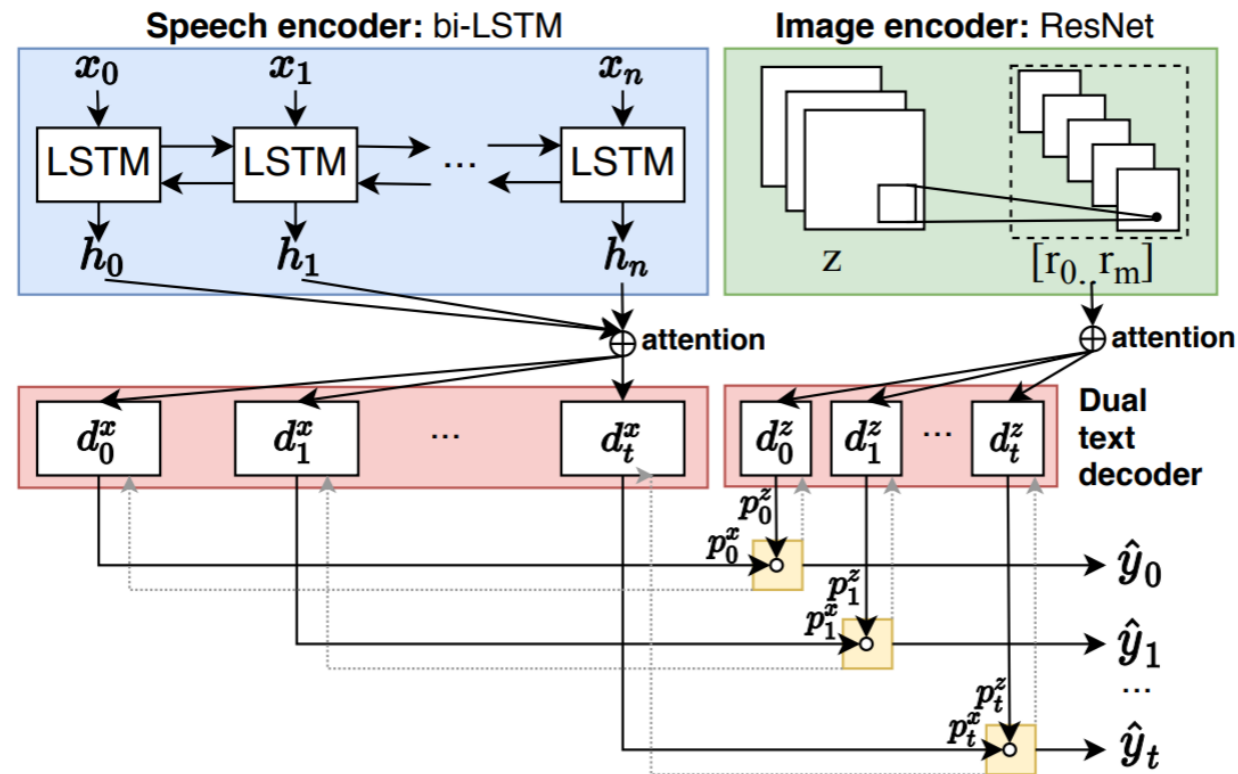
Models used and its evaluation metrics



- ASR: Listen, Attend, and Spell (Chan+, 2016)
 - LSTM encoder-decoder ASR model
 - CER
- TTS: Tacotron (Wang+, 2017)
 - encoder-decoder speech synthesis model
 - using speaker embedding from DeepSpeaker (Li+, 2017) with size of 64
 - L2-norm²
- IC: Show, Attend, and Tell (Xu+, 2015)
 - modified to process 128x128
 - BLEU
- IG: AttnGAN (Xu+, 2017)
 - multistep image generation using adversarial loss
 - generate only until 128x128 image instead of 256x256
 - Inception Score

Multimodal Chain Components

- ImgSp2Txt: average of p_t^x and p_t^z output layer probability
- When only image or speech are available, the decoder uses only the corresponding output layer.
- Trained in character-level granularity to match the best practice of ASR



Experiment Set-up



- Flickr8k + audio (Rashtchian et al, 2010; Harwath and Glass, 2015)
 - 8000 photos of everyday activities and events
 - 5 captions per image, 8920 words vocabulary
 - Crowdsourced natural speech, 183 speakers, 64 hours
- We used the predefined train, dev, test subset
- But the target is to see how the proposed method perform in a single modality dataset
- So we make these **data partition**

Type	Speech	Text	Image	# Image
Multimodal Paired	○	○	○	800
Multimodal Unpaired	Δ	Δ	Δ	1500
Speech only	Δ	x	x	1850
Image only	x	x	Δ	1850

○ : available paired

Δ : available but unpaired

x : unavailable

Result - Comparing MMC1 and MMC2



Training	Data Type	#Image	ASR (CER) ↓	IC (BLEU4) ↑	TTS (L2 ² Norm) ↓	IG (Inception) ↑
MMC 1 Dual-loop (Semi-supervised)	Multimodal (P)	800	36.35	12.75	0.77	5.90
	+ Multimodal (U)	1500	15.10	13.22	0.59	8.29
	+ Sp only (U)	1850	12.37	13.28	0.56	9.12
	+ Img only (U)	1850	12.06	13.29	0.56	9.11
Topline MMC1 (Supervised)	Multimodal (P)	6000	5.76	19.91	0.50	9.66
MMC 2 Single-loop (Semi-supervised)	Multimodal (P)	800	26.67	32.23	0.77	5.90
	+ Multimodal (U)	1500	14.88	55.15	0.65	10.12
	+ Sp only (U)	1850	13.81	58.03	0.62	10.65
	+ Img only (U)	1850	12.32	59.66	0.61	9.95
Topline MMC2 (Supervised)	Multimodal (P)	6000	5.16	79.88	0.50	9.66

ASR improvement even without speech and text data

Discussion – Comparing MMC1 and MMC2



- ASR improvement even when using image dataset
- Sharing between ASR and IC in MMC2 yields better ASR in low-data scenario
- MMC2 ends up with 12.32 CER, on par with MMC1 12.06 CER

- Best score of MMC1 12.06 CER = 17.84 WER
- Comparable with Sun et al. (2016) = 13.81 WER
 - Fully supervised
 - Lattice rescoring algorithm
 - ASR implemented using non end-to-end method, with image to help decoding

Conclusion and Future Works



- Improvements from previous multimodal machine chain:
 - adversarial-based image generation model
 - one-shot speaker adaptation
 - tested on multispeaker natural speech dataset
- Alternative single-loop multimodal chain
- Results shows that both multimodal chains:
 - enables improvement of speech processing components using an image-only dataset
 - by collaborating with image processing components
 - within multimodal machine chain architecture
- Future work: investigate various approaches of component combination

References



- P. Denes and E. Pinson, The Speech Chain, ser. Anchor books. Worth Publishers, 1993.
- A. Tjandra, S. Sakti, and S. Nakamura, “Machine speech chain,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 976–989, 2020.
- —, “Listening while speaking: Speech chain by deep learning,” in Proc. of the IEEE ASRU, Dec 2017, pp. 301–308.
- —, “Machine speech chain with one-shot speaker adaptation,” in Proc. of INTERSPEECH, 2018, pp. 887–891.
- —, “End-to-end feedback loss in speech chain framework via straight-through estimator,” in Proc. of IEEE ICASSP, 2019, pp. 6281–6285.
- J. Effendi, A. Tjandra, S. Sakti, and S. Nakamura, “Listening while speaking and visualizing: Improving ASR through multimodal chain,” in Proc. of the 2019 IEEE ASRU, 2019.
- W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. of IEEE ICASSP, 2016, pp. 4960–4964.
- Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, “Tacotron: Towards end-to-end speech synthesis,” in Proc. of the INTERSPEECH, 2017, pp. 4006–4010.
- K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. of ICML, 2015, pp. 2048–2057.
- T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks,” in Proc. of CVPR, 2018.
- K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. of IEEE CVPR, 2016, pp. 770–778.
- C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, “Deep speaker: an end-to-end neural speaker embedding system,” CoRR, vol. abs/1705.02304, 2017.
- 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, 2010, pp. 139–147.
- D. Harwath and J. Glass, “Deep multimodal semantic embeddings for speech and images,” in Proc. of IEEE ASRU, 2015, pp. 237–244.
- F. Sun, D. Harwath, and J. Glass, “Look, listen, and decode: Multimodal speech recognition with images,” in Proc. of IEEE SLT, Dec 2016, pp. 573–578.
- X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” in Tech. Rep., 2002.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in Proc. of the 40th annual meeting on association for computational linguistics, 2002, pp. 311–318.
- T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in Proc. of the advances in neural information processing systems (NIPS), 2016, pp. 2234–2242.
- Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the bold effect. NeuroImage, 14(2), 427–438. doi: 10.1006/nimg.2001.0812

Thank you for your attention

