LISTENING WHILE SPEAKING AND VISUALIZING: IMPROVING ASR THROUGH MULTIMODAL CHAIN

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

[1]Nara Institute of Science and Technology, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

ABSTRACT

Previously, a machine speech chain, which is based on sequence-to-sequence deep learning, was proposed to mimic speech perception and production behavior. Such chains separately processed listening and speaking by automatic speech recognition (ASR) and text-to-speech synthesis (TTS) and simultaneously enabled them to teach each other in semi-supervised learning when they received unpaired data. Unfortunately, this speech chain study is limited to speech and textual modalities. In fact, natural communication is actually multimodal and involves both auditory and visual sensory systems. Although the said speech chain reduces the requirement of having a full amount of paired data, in this case we still need a large amount of unpaired data. In this research, we take a further step and construct a multimodal chain and design a closely knit chain architecture that combines ASR, TTS, image captioning, and image production models into a single framework. The framework allows the training of each component without requiring a large number of parallel multimodal data. Our experimental results also show that an ASR can be further trained without speech and text data and cross-modal data augmentation remains possible through our proposed chain, which improves the ASR performance.

Index Terms— speech recognition, semi-supervised, multimodal

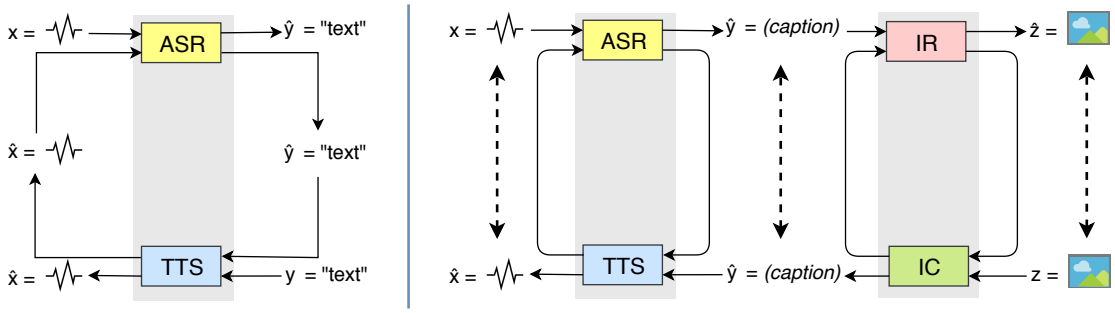
1. INTRODUCTION

Researchers have been working in speech recognition technology for many decades. State-of-the-art ASR systems are currently based on end-to-end deep learning frameworks. Traditionally, they are usually trained by applying supervised learning techniques that rely on the availability of speech data and corresponding transcriptions. To improve the performance in the presence of unexpected acoustic variability, we usually collect more data to train more detailed models. Although some systems have successfully reached parity with humans [1, 2], such a learning style can only be perfectly fit for recognizing the speech of about 10-20 of the world's most common languages. For many others, the problem is the size of the required speech and corresponding transcriptions are usually unavailable.

Recently, approaches that utilize learning from source-to-target and vice-versa as well as feedback links have gained attention and provide the possibility of training models with unpaired datasets. He et al. [3] and Cheng et al. [4], recently published work that addressed a mechanism called dual learning in neural machine translation (NMT). Their system has a dual task: source-to-target language translation (primal) versus target-to-source language translation (dual). It can leverage monolingual data to improve neural machine translation. In image processing, several methods have also been proposed to achieve unsupervised joint distribution matching without any paired data, including DiscoGAN [5], CycleGAN [6], and DualGAN [7]. The framework provides learning to translate an image from a source domain to a target domain without paired examples based on a cycle-consistent adversarial network. Implementation on voice conversion applications has also been investigated [8]. However, most only work with the same domain between the source and the target.

In speech-language processing, the speech chain framework [9, 10, 11] was proposed to integrate human speech perception and production behaviors that utilize the primal model (ASR) that transcribes a text, given the speech versus the dual model (TTS) that synthesizes the speech given the text. Perhaps this is the first framework that was constructed on a different domain (speech versus text). The approach provides freedom from needing a large amount of speech-text paired data and possibilities to improve ASR performance in semi-supervised learning by allowing ASR and TTS to teach each other, given only text or only speech data. Unfortunately, although this approach reduces the requirement of a large amount of paired data, we still need a large amount of unpaired data. Furthermore, this study is limited to speech and textual modalities. In fact, natural communication is actually multimodal that involves both auditory and visual sensory systems.

In this research, we constructed the first framework that accommodates triangle modality (speech, text, and image) and addressed the problems of speech-to-text, text-to-speech, text-to-image, and image-to-text. Our new framework mimics the mechanism of the entire human communication system with auditory and visual sensors. Similar to a machine speech chain, it allows each component to be trained without need-



(a) (b)

Fig. 1. Architecture: (a) speech chain framework [9], and (b) our proposed multimodal chain mechanism.

ing a large number of parallel multimodal data. In addition to the above advantages, through closely knit chain architecture that combines ASR, TTS, IC, and IR models into a single framework, it further frees us from needing a large amount of unpaired data. For example, specifically to ASR tasks, even when no more speech and text data are available, it is still possible to improve ASR by cross-modal data augmentation through our proposed chain.

2. MULTIMODAL CHAIN FRAMEWORK

Figure 1 illustrates (a) the original speech chain framework [9] and (b) our proposed multimodal chain mechanism. In this extension, we included image captioning (IC) and image retrieval (IR) models to incorporate visual modality into the chain. The framework consists of dual loop mechanisms between the speech and visual chains that involve quadruple learning components: ASR, TTS, IC, and IR. In the speech chain, sequence-to-sequence ASR and sequence-to-sequence TTS models are jointly trained in a loop connection, and in the visual chain, neural image captioning and neural embedding-based image retrieval models are also jointly trained in a loop connection. Both chains (speech and visual components) are allowed to collaborate by text modality.

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(x)-text(y)-image(z) data exist: separately train ASR, TTS, IC and IR (supervised learning)**
   Given complete multimodal dataset $\mathcal{D}^{P_{xyz}}$, we can set-up speech utterances x and corresponding text transcriptions y as dataset $\mathcal{D}_{xy}^{P_{xyz}}$ to separately train both the ASR and TTS models in a supervised manner. ASR losses L_{ASR}^{P} and L_{TTS}^{P} are calculated directly with teacher-forcing, where the ground truth for each time step is used as input when decoding. We can also set-up images z and captions y as dataset $\mathcal{D}_{yz}^{P_{xyz}}$ to separately train the IC and IR models with supervised learning. The IC model is trained with teacher-forcing on reference caption y, and the IR model is trained with pairwise rank loss [12] on reference image z and its contrastive sample.

2. **Unpaired speech (x), text (y), images (z) data exist: jointly Train ASR&TTS in the speech chain and IC&IR in the visual chain (unsupervised learning)**
   In this case, although speech, text, and image data are available, they are unpaired.

   (a) **Only using speech data: unrolled process ASR→TTS in a speech chain**
   Here we only use speech utterances x of dataset $\mathcal{D}^{U_{xyz}}$, and ASR generates transcription $\hat{y}$ for TTS to reconstruct. Reconstructed transcriptions $\hat{x}$ calculate loss L_{TTS}^{U} between x and $\hat{x}$ and update the model parameter.

   (b) **Only using image data: unrolled process IC→IR in a visual chain**
   Using only image z in dataset $\mathcal{D}^{U_{xyz}}$, image captions $\hat{y}$ are generated with the IC model. These captions are then used by the IR model to update its multimodal space using pairwise rank loss, which resulted in loss L_{IR}^{U}.

   (c) **Only using text data: unrolled process TTS→ASR in the speech chain and IR→IC in the visual chain**
   Given only the text in dataset $\mathcal{D}^{U_{xyz}}$, TTS generates speech utterance $\hat{x}$ for the ASR, which then reconstructs the speech utterances into text $\hat{y}$ in which reconstruction loss L_{ASR}^{U} between y and $\hat{y}$ can be calculated. On the other hand, image captions y retrieve images $\hat{z}$, which are reconstructed into text $\hat{y}$ using the IC model in which losses L_{IC}^{U} are calculated between y and $\hat{y}$.

3. **Single data (either speech (x), text (y), or images (z)) exist: train ASR & TTS jointly in the speech chain and IC & IR in the visual chain (unsupervised learning)**
   In this case, only a single modality data (either speech, text, or image) is available, and the others are empty.

   (a) **Only text data exist: train the speech and visual chains, as in 2(c)**

If only text data are available in dataset $\mathcal{D}^{U_y}$, we can separately perform unrolled process TTS→ASR in the speech chain and IR→IC in the visual chain

(b) **Only speech data exist: speech chain → visual chain**
If only speech data are available in dataset $\mathcal{D}^{U_x}$, first we perform unrolled process ASR→TTS in the speech chain (See 2(a)). The generated text transcription $\hat{y}$ is then used to performs IR→IC in the visual chain (See 2(c)).

(c) **Only image data exist: visual chain → speech chain**
If only image data are available in dataset $\mathcal{D}^{U_z}$, 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)).

Here our main concern is the last point (3(c)). We are interested to learn whether in the situation where only image data exist (no more speech and text data are available) we can still improve the ASR performance through a learning process from a visual chain to a speech chain by leveraging cross-modal data augmentation.

We combine all of the losses and update both the ASR and TTS models as well as the IC and IR models:

$$L_{sc} = \alpha_{ASR} L^P_{ASR} + \alpha_{TTS} L^P_{TTS} + \beta_{ASR} L^U_{ASR} + \beta_{TTS} L^U_{TTS} \quad (1)$$

$$\theta_{ASR} = Optim(\theta_{ASR}, \nabla_{\theta_{ASR}} L) \quad (2)$$

$$\theta_{TTS} = Optim(\theta_{TTS}, \nabla_{\theta_{TTS}} L) \quad (3)$$

$$L_{vc} = \gamma_{IC} L^P_{IC} + \gamma_{IR} L^P_{IR} + \delta_{IC} L^U_{IC} + \delta_{IR} L^U_{IR} \quad (4)$$

$$\theta_{IC} = Optim(\theta_{IC}, \nabla_{\theta_{IC}} L) \quad (5)$$

$$\theta_{IR} = Optim(\theta_{IR}, \nabla_{\theta_{IR}} L) \quad (6)$$

which results in losses L_{sc} and L_{vc} for the speech and visual chains. Parameters α, β, γ, and δ are hyper-parameters for scaling the loss between the supervised (paired) and unsupervised (unpaired) losses in each chain.

3. MULTIMODAL CHAIN COMPONENTS

In this section, we briefly describe all of the components inside the multimodal chain framework.

3.1. Sequence-to-sequence ASR

We use the sequence-to-sequence ASR model with attention, whose architecture resembles a previous one used in [9] that is also based on LAS framework [13]. It directly models conditional probability $P(y|x)$ of transcription y given speech feature x. For the speech feature, the MFCC or the mel-spectogram are usually encoded by a bidirectional LSTM encoder. The hidden representations are then attended by a LSTM or GRU decoder that decodes a sequence of characters or phonemes.

3.2. Sequence-to-sequence TTS

A sequence-to-sequence TTS is a parametric TTS that generates sequence of speech feature x from transcription y. We also used similar architecture as a previous one used in [9] which is based on a slight modification to Tacotron [14]. Tacotron produces a mel-spectogram given the text utterances, and is further transformed into a linear spectogram so that the speech signal can be reconstructed using the Griffin-Lim algorithm [15].

3.3. Image Captioning

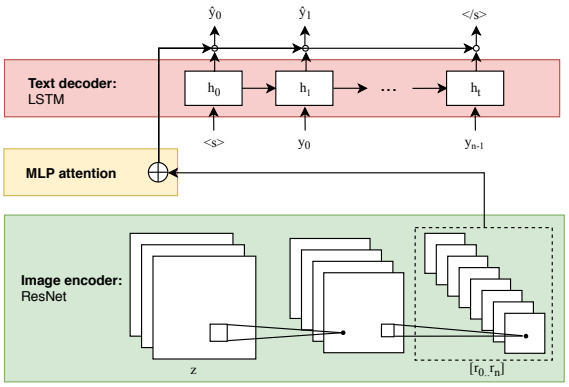


Fig. 2. Caption model

An image captioning model receives image z and produces caption prediction $\hat{y} = [\hat{y}_0..\hat{y}_n]$ by minimizing softmax cross entropy loss against original caption $y = [y_0..y_n]$. We utilized similar architecture as [16], where image z are encoded through a series of convolutional neural network enc_{img} resulting in a high level feature representation within a certain number of region $enc_{img}(z) = [r_0..r_n]$ that represent parts of the image. During decoding the $[\hat{y}_0..\hat{y}_n]$, linear attention grounds each decoded word into correlated image region r_n by calculating alignment probability $a_t(enc_{img}(z)) = Align([r_0..r_n], h_t)$ over decoder states h_t. Unlike Xu et al's model, we use ResNet [17] instead of VGG [18].

3.4. Image Retrieval

Neural IR models [19, 20, 21, 22] are implemented by realizing a multimodal embedding between image z and its caption y. Image embedding z^e is usually extracted from a series of pretrained convolutional neural networks followed by linear transformation. Recurrent neural network encoder are used to generate caption embedding y^e, which is transformed from the last encoder state h_n that encodes sequence of words $[y_0..y_n]$. These two embedding representations are then trained using pairwise rank loss function L_{IR} to join them into a unique multimodal embedding space.

As shown in Eq. 7, this procedure reduces mean squared distance d between each image embedding z^e with related

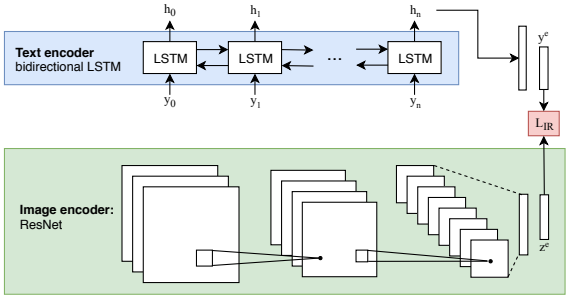


Fig. 3. Retrieval model

caption represented with text embedding y^e, and increases its distance with unrelated caption $\hat{y}^e$. Margin M is used to distance the already similar pairs, providing space to optimize the hard-positive examples:

$$L_{IR} = \sum_{|y^e|}\sum_{|\hat{z}^e|} max\{0, M - d(y^e, z^e) + d(y^e, \hat{z}^e)\} + \sum_{|z^e|}\sum_{|\hat{y}^e|} max\{0, M - d(z^e, y^e) + d(z^e, \hat{y}^e)\} \quad (7)$$

4. EXPERIMENTAL SET-UP

4.1. Corpus Dataset

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 ×.

Dataset	Speech x	Text y	Image z	# Data*	Training Type
$D^{P_{xyz}}$	◯	◯	◯	2000	1 (paired)
$D^{U_{xyz}}$	▲	▲	▲	7000	2 (unpaired)
D^{U_x}	▲	×	×	10000	3 (unpaired)
D^{U_z}	×	×	▲	10000	3 (unpaired)

We ran our experiment with the Flickr30k dataset [23] that has 31,014 photos of everyday activities, events, and scenes. Similar to other image captioning datasets, each image has five captions with a vocabulary of 18k words. However, since we use this dataset not only for captioning but also for retrieval, we need to maintain a balance between source and target. For image captioning, we use one caption per image to make the learning target consistent by avoiding one-to-many confusion. Conversely, in image retrieval we used all five captions because the learning target is already consistent. To train the speech counterpart of our proposed architecture, we generated speech from the Flickr30k captions using single speaker Google TTS. The transcription resulted in 145k utterances, all with the same speaker, with the total duration of 178.8 hours. Each caption typically consists of 12.32 words in the form of a sentence.

In our experiment, we used the dev and test set of Flickr30k for validation and testing, respectively. Similar to the training data, Google TTS is also used to generate the corresponding speech (single speaker voice) of these two sets.

To show the capability of our model for semi-supervised learning, we formulated the training part of the dataset into four parts. For more details, see the specifications in Section 2. The selection of which data belongs to which part was done randomly since all the training data are shuffled before being split into four parts. The first part, $D^{P_{xyz}}$, was used to supervisedly train each model (**Type 1**), because all the data are paired (◯). Next, the $D^{U_{xyz}}$ dataset, which has all the modalities available but unpaired (▲) was used to separately train the speech and visual chains (**Type 2**). D^{U_x} and D^{U_z} are assumed to be a single modality corpus (×), which only has speech and images without any transcription or captioning. By decoding the D^{U_x} dataset into image captions, and D^{U_z} into utterance transcriptions, we can use the generated data to further semi-supervisedly improve each model (**Type 3**).

Without our proposed architecture, these monomodal data D^{U_x} and D^{U_z} cannot be used because their modality is completely unrelated to the chain in the other modality pair. As mentioned above, our main concern is to know whether it is still possible to improve ASR performance in the **Type 3** situation, where only image data is available on D^{U_z}.

4.2. Model Details

We used a standard sequence-to-sequence model with an MLP attention module for ASR as mentioned in Section 3.1. For the TTS, we used Tacotron 3.2. The features and architecture of the ASR and TTS models are similar with Tjandra et al. (2017) single speaker speech chain models [9]. We used an Adam optimizer with a learning rate of 1e-3 for the ASR model, 2.5e-4 for the TTS model, and 1e-4 for the IC model. For the IR model, we used a stochastic gradient descent with a 0.1 learning rate.

In the visual chain, we implemented the IC and IR models as previously described in Sections 3.3 and 3.4. For the convolutional part that extracts the image features, we used ResNet [17] as an image encoder in IC and IR. In the IC model, we removed the last two layers of the ResNet, which yields a 14x14 latent representation of the image region in which the decoder could attend to. Then, for the IR model, we removed the last layer, giving us a 2048-dimensional hidden representation that can be regarded as an image representation. These representations are then linearly transformed into 300-dimensional image embeddings. On the other hand, we generated text embeddings using a single-layer bidirectional LSTM with 256 hidden sizes in each direction.

We decoded the transcription in the speech chain using beam-search decoding with a size of three. Similarly, during the visual chain operation, the IC model produced its hypoth-

*Different modality has different unit (speech=utterance, text=sentence, image=picture). In paired data, one image is at least associated with one text sentence and one speech utterance.

esis using beam-search decoding of the same size. The granularity difference between each chain (i.e., char and word) was fixed during training between the chains. To simulate sampling in the IR hypothesis, we randomly sampled one hypothesis from five candidates.

5. EXPERIMENT RESULTS

5.1. A Large Amount of Paired Data Exists - Topline Case

Table 2. Our ASR and TTS performance in comparison with the existing published results

Data	ASR CER(%)	TTS L2-norm^2
Kim et al. [24]	11.08	-
Tjandra et al. [25]	6.60	0.682
Ours	6.87	0.653

Table 3. Our IC and IR performance in comparison with the existing published results

Data	IC	IR	
	BLEU1	R@10↑	med r↓
Xu et al. (2015) [16]	67.00	-	-
Vilalta et al. (2017) [20]	-	59.8	6
Ours	66.27	62.42	5

In this subsection, we assumed that a large amount of paired data exists. Therefore, we can train each model independently using supervised training. Here we compared the performance of all the system components with the existing published results on a well-known dataset. For the ASR and TTS tasks, we evaluated the performance of our models on the Wall Street Journal dataset [26], which is a natural multispeaker speech corpus. Our settings for the training, development, and test sets are identical as the Kaldi s5 recipe [27]. We trained our model with the WSJ-SI284 data. Our validation set was dev93, and our test set was eval92. For the IC and IR tasks, we evaluated the performance of our models on a full set of Flickr30k.

Tables 2 and 3 show a comparable evaluation for the ASR-TTS and IC-IR tasks. As seen in the table, our ASR performs better than Kim et al. [24] and provides a similar performance to Tjandra et al. [10]. Our TTS model also performs on par with Tjandra et al. [10]. Our IC model also performed on par on the Flickr30k dataset with the work by Xu et al. by a 0.7 BLEU1 margin. Finally, we compared our IR model with Vilalta et al. (2017) who proposed a full network embedding model in which our model performed slightly better. These results reveal that in a fully supervised scenario, our model works as well as previously published papers.

*We trained our baseline model with the only 2k paired data to simulate a real-condition where an only small amount of paired dataset is available to show that our chain can improve the initial model that was only trained with a small amount of data using semi-supervised learning fashion.

Table 4. ASR and TTS performance using Multimodal Chain

Data	ASR WER(%)	TTS L2-norm^2
Baseline: ASR & TTS (Supervised learning - Type 1)		
$D_{xy}^{P_{xyz}}$ 2k*	81.31	0.874
Proposed: speech chain ASR→TTS and TTS→ASR (Semi-supervised learning - Type 2(a)&2(c))		
$+D_{xy}^{U_{xyz}}$ 7k	10.60	0.714
Proposed: visual chain → speech chain Semi-supervised learning - Type 3(b)&3(a))		
$+D^{U_z}$ 10k	7.97	0.645
Topline: ASR & TTS separately (Supervised learning - Full Data)		
$D_{xy}^{P_{xyz}}$ 29k	2.37	0.398

Table 5. IC and IR performance using Multimodal Chain

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

5.2. Only a Small Amount of Paired Data Exists - Utilizing Multimodal Chain

As explained in Sec. 2, in this section we demonstrate how our proposed multimodal chain mechanism deal with the data composition written in Table 4.1 where there are only 2k speech-text-image paired data, 7k speech-text-image unpaired data and 10k single modality data.

Note that in this research, the idea is not to show that the multimodal chain can outperform a baseline that was only trained with a small dataset. Instead, we want to identify how much we can improve performance when only unpaired data are available or even how much more improvement is possible when the required data are no longer available. We are interested to see how much we can improve the ASR performance when no more speech and text are available to train the model.

Table 4 shows the ASR and TTS results from the scenarios in Section 2. First, we trained them on 2k paired data $D_{xy}^{P_{xyz}}$ as shown in the first block using the supervised train-

ing method. This initial model is then used in the next step as the speech chain component. We continued the training into the speech chain using $D_{xy}^{U_{xyz}}$ 7k data and achieved 10.60% WER and 0.714 L2-norm^2. Finally, using the IC model that was trained semi-supervisedly through Type 2(a)&2(c), we decoded the image-only D^{U_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 [9] that was only trained using the speech and text datasets. This result proved that the cross-modal data augmentation from the image modality into this speech chain is correlated positively with model quality. Our proposed strategies makes improvement of ASR and TTS possible, even without any speech or text data, with the help of a visual chain.

Table 5 shows the IC and IR results from similar scenarios with the improvement from the speech chain. First, we did training using paired 2k data and achieved the baseline score shown in the first block. Next, we semi-supervisedly trained the IC and IR models in the visual chain mechanism, and produced over 8.2 BLEU1 improvement, 1.26 recall at 10 (R@10) improvement and 3 point improvement for the median r metrics. Finally, in the third block 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.

6. RELATED WORKS

Human communication is multisensory and involves several communication channels, including auditory and visual channels. Such multiple signals in different form are processed based on their characteristics, but later can be used for mutual augmentation. Moreover, the sensory inputs from several modalities share complementary behavior to ensure a robust perception of the overall information.

Over the past decades, several studies have integrated audio and visual cues to improve speech recognition performance. Within recent deep learning frameworks, Petridis et al. [28] proposed one of the first end-to-end audiovisual speech recognition schemes. Another approach is the "Watch, Listen, Attend, and Spell (WLAS)" framework [29], which is an extension of the LAS framework [13] for speech recognition tasks that utilize a dual attention mechanism that can operate in three ways: over visual input only, audio input only, or both. Afouras et al. [30] also proposed a deep audio-visual speech recognition to recognize phrases and sentences spoken by a talking face with or without audio.

Therefore, although the idea of incorporating visual information for automatic speech recognition (ASR) is basically not new, most approaches are usually done by simply concatenating the information that is inefficient to effectively fuse information from various modalities. Furthermore, these methods require that all of the information from these modalities be available altogether, which is often difficult. In contrast, humans process different modalities by different organs (i.e., ears for listening, mouths for speaking, eyes for seeing, etc.), which enables independent processing of such information while simultaneously opening room for augmenting each other.

Previously, a machine speech chain, based on sequence-to-sequence deep learning, was proposed to mimic speech perception and production behavior. It separately processed listening and speaking by ASR and a text-to-speech synthesis (TTS), but also enabled semi-supervised learning to teach each other when they received unpaired data. Unfortunately, our study is limited to speech and textual modalities. In fact, natural communication is actually multimodal because it involves both auditory and visual sensory systems

In this research, we take a further step to construct a multimodal chain and design a closely knit chain architecture that combines ASR, TTS, image captioning, and image production (retrieval or generation) models into a single framework. The framework allows each component to be trained without a large number of parallel multimodal data. Our experimental results show that further training of an ASR without available speech and text data remains possible by cross-modal data augmentation through our proposed multimodal chain, which improves ASR performance.

7. CONCLUSION

We described a novel approach for cross-modal data augmentation that upgrades a speech chain into a multimodal chain. We proposed a visual chain by jointly training IC and IR models in a loop connection that can learn semi-supervisedly over an unpaired image-text dataset. Then 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. We showed that each model in both chains can assist each other given an incomplete dataset by leveraging the data augmentation among modalities. In the future, we will jointly train both the speech and visual chain so that both can also be updated together. Furthermore, following the previous speech chain [10] that can synthesize multi-speaker speech, we will investigate our multimodal chain on natural multispeaker speech dataset as well.

8. ACKNOWLEDGEMENTS

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.

9. REFERENCES

[1] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5934–5938.

[2] 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*, 2017, pp. 132–136.

[3] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Advances in Neural Information Processing Systems*, 2016, pp. 820–828.

[4] Y. Cheng, Z. Tu, F. Meng, J. Zhai, and Y. Liu, "Towards robust neural machine translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1756–1766.

[5] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 1857–1865.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[7] Z. Yi, H. R. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 2868–2876.

[8] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, 2018, pp. 632–639.

[9] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," *CoRR*, vol. abs/1707.04879, 2017. [Online]. Available: http://arxiv.org/abs/1707.04879

[10] ——, "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.*, 2018, pp. 887–891. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1558

[11] ——, "End-to-end feedback loss in speech chain framework via straight-through estimator," *CoRR*, vol. abs/1810.13107, 2018.

[12] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, ser. IJCAI'11. AAAI Press, 2011, pp. 2764–2770. [Online]. Available: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-460

[13] W. Chan, N. Jaitly, Q. Le, and O. 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)*. IEEE, 2016, pp. 4960–4964.

[14] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," 2017.

[15] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, April 1984.

[16] 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 *International conference on machine learning*, 2015, pp. 2048–2057.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[19] J. Dong, X. Li, and C. G. M. Snoek, "Word2visualvec: Cross-media retrieval by visual feature prediction," *CoRR*, vol. abs/1604.06838, 2016. [Online]. Available: http://arxiv.org/abs/1604.06838

[20] A. Vilalta, D. Garcia-Gasulla, F. Parés, E. Ayguadé, J. Labarta, E. U. Moya-Sánchez, and U. Cortés, "Studying the impact of the full-network embedding on multimodal pipelines," *Semantic Web*, no. Preprint, pp. 1–15.

[21] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2623–2631, 2015.

[22] I. Calixto and Q. Liu, "Sentence-level multilingual multi-modal embedding for natural language processing," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., Sep. 2017, pp. 139–148.

[23] 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)*, Dec 2015, pp. 2641–2649.

[24] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[25] A. Tjandra, S. Sakti, and S. Nakamura, "Multi-scale alignment and contextual history for attention mechanism in sequence-to-sequence model," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 648–655.

[26] 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*. Association for Computational Linguistics, 1992, pp. 357–362.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[28] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end audiovisual fusion with lstms," in *Auditory-Visual Speech Processing, AVSP 2017, Stockholm, Sweden, 25-26 August 2017.*, 2017, pp. 36–40.

[29] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[30] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.