IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 12, 2021, accepted April 22, 2021, date of publication May 6, 2021, date of current version May 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077886

Multimodal Chain: Cross-Modal Collaboration Through Listening, Speaking, and Visualizing

JOHANES EFFENDI[1,2], ANDROS TJANDRA[1], SAKRIANI SAKTI[1,2], (Member, IEEE), AND SATOSHI NAKAMURA[1,2], (Fellow, IEEE)

[1]Nara Institute of Science and Technology, Ikoma 630-0192, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Tokyo 103-0027, Japan

Corresponding author: Sakriani Sakti (ssakti@is.naist.jp)

Part of this work is supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers JP17H06101 and JP21H03467, as well as Google AI Focused Research Awards Program.

ABSTRACT Language is an integral part of human interpersonal communication, which is conveyed through multiple sensory channels. This multisensory communication skill has motivated an extensive number of studies on multimodal information processing, which are trying to develop a system that mimics this natural behaviour. For example, automatic speech recognition (ASR) represents listening activity, text-to-speech (TTS) represents speaking, and various image processing models to represent visual perception. Most are trained and tuned independently using parallel examples from the source to the target modality. However, this is not the case in real-life situations, where a lot of paired data are unavailable. Inspired by this self-supervision of the human auditory and visual perception system, we proposed a multimodal chain mechanism with a weakly-supervised chain training strategy that is trained and tuned jointly. In our proposed framework, when the amount of paired training data are insufficient, collaboration among ASR and TTS, image captioning (IC), and image production models can improve their performance through single or dual-loop chain mechanisms. Our experiment result showed that by using such a closed-loop chain mechanism, we can improve a model with both unpaired and unrelated data from different modalities in a semi-supervised manner. Through the collaboration of speech and visual chains, we improve an ASR model performance with an image-only dataset while maintaining the performance of other models.

INDEX TERMS Semi-supervised learning, multimodal machine chain, automatic speech recognition, speech chain.

I. INTRODUCTION

Humans perceive the world through different modalities which they process with their senses. Such multiple information from multiple modalities is processed altogether into a general concept and understanding. In a work on human listening and speaking activities, Denes *et al.* described how closely related they are to each other, even though these activities are done by different organs with different purposes. They called this mechanism a speech chain [1], where spoken messages are propagated from the speaker's mind to the listener's mind (Figure 1, left). During the speech production process, the hearing process is not only needed by the interlocutor but also by the speaker. Through simultaneous speaking and listening, the speaker can monitor her speech quality with self-supervision from her brain.

Inspired by this speech chain mechanism (Figure 1), we previously proposed a machine speech chain [2]–[5] by exploiting the relation between human speech perception and production. This approach enables the training of ASR and text-to-speech synthesis (TTS) with speech-only or text-only data. First, both the ASR and TTS models are trained on a small amount of paired speech-text data. Then both models generate pseudo-pairs of the unpaired data in an online manner. In this step, the chain mechanism relies on reconstruction loss, where the only supervision comes from comparing the chain hypothesis with the original unpaired data.

Although this speech chain approach involves ASR and TTS tasks, its improvement still comes from the data whose modality is related to the task. We can only use speech and text data to improve ASR and TTS, which are the source

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

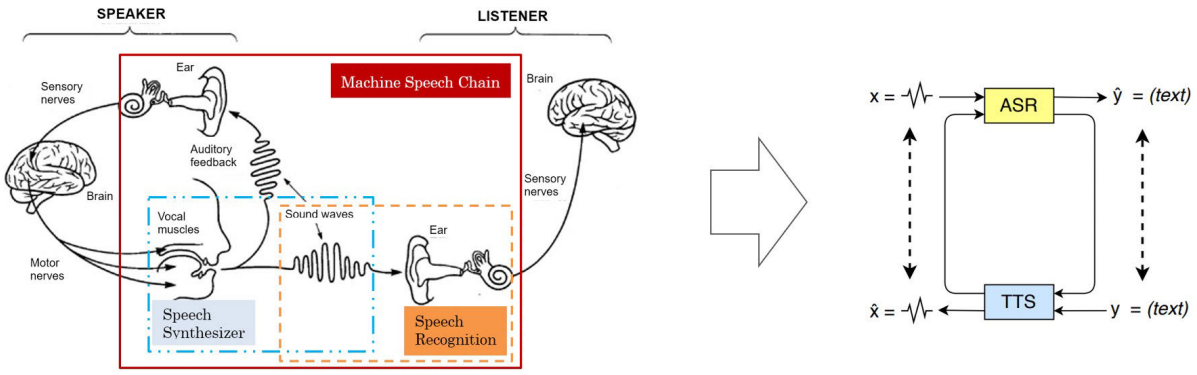


FIGURE 1. Human speech chain [1] in comparison with machine speech chain [2].

and target modalities of each task. From the perspective of the human cognitive process, human communication is multimodal, where each modality shares complementary behavior to ensure flexible learning. When it is impossible to infer from auditory channels, a visual portion can complement the missing information [6]. In addition to serving as a backup, both modalities can also be perceived together, reflected in evidence of cross-modal processing in the brain [7]. Byrne (2010) argued that vision supplies information about color, texture, and other visual aspects of viewed objects, and the other senses supply different features to help us perceive what we see [8].

Inspired by this idea, we should be able to improve our speaking ability (i.e., speech processing models) with different kinds of stimuli (data), such as visual stimuli (i.e., image data). In this study, we focus on this generalization of the speech chain into a multimodal chain by improving the speech processing models by leveraging visual data. Moreover, using such visual data as images is beneficial because unlabeled images are available in practically unlimited quantities [9]. Inside the multimodal chain, the ASR and TTS models in the speech chain collaborate with IC and image production models (image retrieval (IR) or image generation (IG)) from the visual chain. In this way, we can leverage cross-modal augmentation inside a chain of models, especially when the paired training data are insufficient.

Our contributions in this study are four-fold.[1] First, we define a general framework for the universal chain problem and formalize the definition of the semi-supervised chain mechanism, making it applicable to any set of modalities. Second, by implementing the general framework, we reformulate the previously published speech chain into a multimodal machine chain. Unlike speech chain that is limited to the related modality (i.e., speech and text for ASR and TTS), we can use unrelated modality data (i.e., image data for speech processing) with our dual-chain mechanism. With this idea, we significantly reduce the need for parallel data for training a speech processing model because the models inside our chain mechanism support each other through a joint training mechanism.

Then, we improve the robustness of our multimodal chain. We experimented with an adversarial-based image generation model for generating unseen images. From speech processing aspects, we added a speaker recognition model to our chain mechanism, which enables the use of a multi-speaker dataset through a one-shot speaker adaptation. We also investigated the performance of our chain mechanism on synthetic speech and on a natural speech dataset.

Finally, we propose an alternative single-loop multimodal chain with an ImgSp2Txt model that receives image and speech input altogether when both are available and then decodes the text transcription. This enables sharing between ASR and IC tasks, which is commonly investigated in the audiovisual ASR field. We want to determine whether our multimodal chain training mechanism can also be applied for such a multi-source multimodal model.

II. GENERAL FRAMEWORK FOR THE CHAIN MECHANISM

In a cross-modal $X \rightarrow Y$ mapping task, we define the source modality as X, target modality as Y, and unrelated modality as Z. Suppose there are three kinds of data based on its availability:

- P_{xyz} is paired $\{X, Y, Z\}$ trimodal data,
- $U_{x,y,z}$ is unpaired data, where there is no mapping between each row of x and each row of y or z,
- and S_z is single modality data, whose modality Z has no relation with the task modality (i.e. X and Y).

In this section, we describe how to train this cross-modal model $M_{X\rightarrow Y}$ based on the data availability.

A. STEP 1: SUPERVISED TRAINING USING PAIRED DATA

Given enough data pairs of $\{(x_0^P, y_0^P), (x_1^P, y_1^P), \ldots, (s_n^P, t_n^P)\} \in P_{xy}$, cross-modal model $M_{X\rightarrow Y}$ can be trained in a supervised manner by minimizing the loss between predicted $\hat{y}_i^P = M_{X\rightarrow Y}(x_i^P)$ and ground truth $y_i^P \in P_{xy}$ so that:

$$\ell_{M_{X\rightarrow Y}} = L_{M_{X\rightarrow Y}}(y_i^P, \hat{y}_i^P; \theta_{M_{X\rightarrow Y}}), \quad (1)$$

$$\theta_{M_{X\rightarrow Y}} = Optim(\theta_{M_{X\rightarrow Y}}, \nabla_{\theta_{M_{X\rightarrow Y}}} \ell). \quad (2)$$

[1] Parts of this work were presented in previous work [10], [11]. This manuscript contains a general framework for the universal chain problem that has not been defined before. Furthermore, we also include additional experiments and detailed analysis.

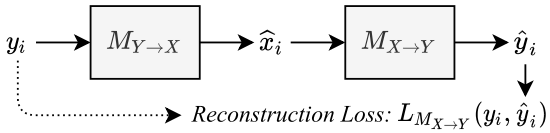


FIGURE 2. Illustration of chain path $\mathcal{C}_{YXY} = \{Y \to X, X \to Y\}$ with $|D| = 2$, where $M_{X\to Y}$ is backpropagated by the reconstruction loss $L_{M_{X\to Y}}$.

B. STEP 2: SEMI-SUPERVISED TRAINING USING UNPAIRED DATA

Since in some cases, there are insufficient data pair in P_{xy} so that $\{(x_0^P, y_0^P), (x_1^P, y_1^P), \ldots, (x_m^P, y_m^P)\} \in P_{xy}, m < n$, cross-modal model $M_{X\to Y}$ cannot be optimally trained to get satisfiable quality. Given unpaired data $\{x_0^U, x_1^U, \ldots, x_n^U\} \in U_x$ and $\{y_0^U, y_1^U, \ldots, y_n^U\} \in U_y$, cross-modal model $M_{X\to Y}$ training can continue to use the chain mechanism by leveraging its inverse model $M_{Y\to X}$.

In this condition, we can construct chain path $\mathcal{C}_{YXY}$ (See Figure 2) to continue training model $M_{X\to Y}$:

$$\mathcal{C}_{YXY} = \{Y \to X, X \to Y\}, \tag{3}$$

by generating hypothesis $\hat{x}_i^U$ from inverse model $M_{Y\to X}$:

$$\hat{x}_i^U = M_{Y\to X}(y_i^U), \tag{4}$$

so that the hypothesis of $\hat{y}_i^U$ can be generated:

$$\hat{y}_i^U = M_{X\to Y}(\hat{x}_i^U), \tag{5}$$

which enables the calculation of reconstruction loss $\ell_{M_{X\to Y}}$:

$$\ell_{M_{X\to Y}} = L_{M_{X\to Y}}(y_i^U, \hat{y}_i^U; \theta_{M_{X\to Y}}). \tag{6}$$

In an end-to-end condition, $M_{X\to Y}$ can be backpropagated:

$$\theta_{M_{X\to Y}} = Optim(\theta_{M_{X\to Y}, M_{Y\to X}}, \nabla_{\theta_{M_{X\to Y}, M_{Y\to X}}} \ell), \tag{7}$$

while in a non end-to-end condition, Eq. 2 is sufficient.

Given the mechanism, the improvement of $M_{X\to Y}$ is dependent on the quality of $\hat{x}_i^U$ which functions as a bridge between $Y \to X$ and $X \to Y$. Therefore, reciprocally training inverse model $M_{X\to Y}$ with inverse chain operation $\mathcal{C}_{XYX}$ is also encouraged.

C. STEP 3: SEMI-SUPERVISED TRAINING USING SINGLE MODALITY DATA

When all paired P_{xy} data and unpaired U_{xy} data have been used, model $M_{X\to Y}$ can still be improved by generalizing the chain mechanism explained in Sec. II-B. This generalization enables the use of unrelated modality $\mathbf{Z}$ to improve model $M_{X\to Y}$ which was previously only trained within $\{X, Y\}$ modalities.

First, let us assume now that we have three kind of modalities $D = X, Y, Z$, and paired data $\{(x_0^P, y_0^P, z_0^P), (x_1^P, y_1^P, z_1^P), \ldots, (x_m^P, y_m^P, z_m^P)\} \in P_{xyz}, m < n$, which are inadequate to satisfiably train $M_{X\to Y}$ as in Sec. II-A. Similar to Sec. II-B, single-modality data $\{x_0^S, x_1^S, \ldots, x_n^S\} \in S_x$, $\{y_0^S, y_1^S, \ldots, y_n^S\} \in S_y$, and $\{z_0^S, z_1^S, \ldots, z_n^S\} \in S_z$ are available. In this condition, we can construct chain path $\mathcal{C}_{ZYXY}$ that leverages S_Z single-modality data:

$$\mathcal{C}_{ZYXY} = \{Z \to Y, Y \to X, X \to Y\}, \tag{8}$$

by generating hypothesis $\hat{y}_i^S$ with model $M_{Z\to Y}$:

$$\hat{y}_i^S = M_{Z\to Y}(z_i^S), \tag{9}$$
$$\hat{x}_i^S = M_{Y\to X}(\hat{y}_i^S), \tag{10}$$
$$\hat{y}_i^S = M_{X\to Y}(\hat{x}_i^S), \tag{11}$$

which enables the calculation of reconstruction loss $\ell_{M_{X\to Y}}$:

$$\ell_{M_{X\to Y}} = L_{M_{X\to Y}}(\hat{y}_i^S, \hat{\hat{y}}_i^S; \theta_{M_{X\to Y}}). \tag{12}$$

In an end-to-end condition, $M_{X\to Y}$ can be backpropagated:

$$\theta_{M_{X\to Y}} = Optim(\theta_{M_{X\to Y}, M_{Y\to X}, M_{Z\to Y}}, \nabla_{\theta_{M_{X\to Y}, M_{Y\to X}, M_{Z\to Y}}} \ell), \tag{13}$$

while in a non end-to-end condition, Eq. 2 is sufficient. Figure 3 illustrates this chain path.

As we can see from the process flow, Eq. 10-13 are similar with Eq. 4-7 because the chain path $\mathcal{C}_{YXY}$ are inside the path of $\mathcal{C}_{ZYXY}$. Therefore, we can develop an extension from the chain with $|D| = 2$ to $|D| = 3$, which further shows the generalization of the chain framework.

III. BASIC MACHINE SPEECH CHAIN

This section describes the machine speech chain [2]–[5] as a chain implementation with two modalities ($|D| = 2$). In this framework, ASR and TTS models are trained in a closed-loop mechanism that allows semi-supervised training using both paired and unpaired speech and text data. We use the definition in Section II to describe the machine speech chain:

- X source modality is speech, Y target modality is text,
- $M_{X\to Y}$ model is ASR, $M_{Y\to X}$ inverse model is TTS,
- both ASR and TTS models are trained with a small amount of P_{xy} paired speech-text data,
- $\mathcal{C}_{YXY}$ is an unsupervised step to improve the ASR model using U_y unpaired text data,
- $\mathcal{C}_{XYX}$ is an unsupervised step to improve the TTS model using U_x unpaired speech data.

In this study, we demonstrate the generalization of this chain mechanism for any number of modalities of $|D| > 2$ by developing a multimodal machine chain with $|D| = 3$.

IV. MULTIMODAL MACHINE CHAIN

We realize the generalization of a semi-supervised chain mechanism when $|D| = 3$ with three kinds of modalities, X, Y, Z for speech, texts, and images.

A. DUAL-LOOP MULTIMODAL CHAIN (MMC1-IR/IG)

To connect each of these modalities in this multimodal chain, we defined five kinds of models and our proposed chain path to improve them in a semi-supervised manner with single-modality data:

- $M_{X\to Y}$ is an automatic speech recognition (ASR) model that transcribes speech (X) into text (Y).
- $M_{Y\to X}$ is a text-to-speech synthesis (TTS) model that synthesizes speech (X) from text (Y),

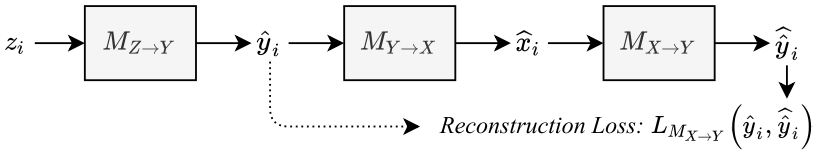


FIGURE 3. Illustration of chain path $\mathcal{C}_{YXY}$ into $\mathcal{C}_{ZYXY} = \{Z \to Y, Y \to X, X \to Y\}$ with $|D| = 3$ for enabling the semi-supervised chain training from single modality data $z_i \in S_z$.

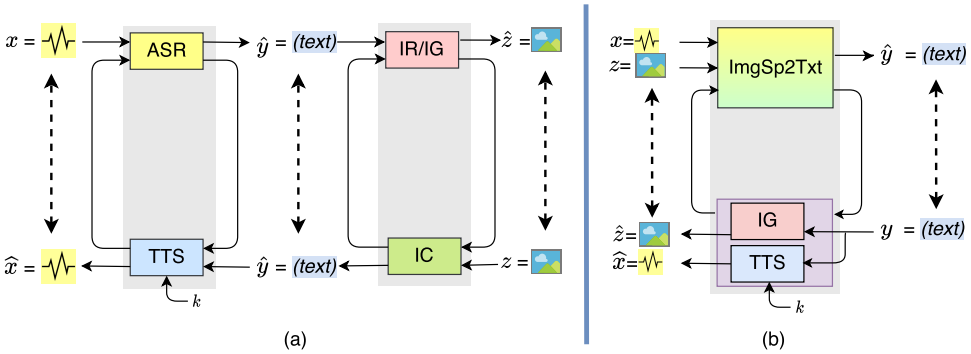


FIGURE 4. Structure of: (a) proposed dual-loop multimodal chain with image retrieval (IR) or image generation (IG) (MMC1-IR/IG), (b) proposed single-loop multimodal chain (MMC2). (Dashed arrow denotes a comparison, and k denotes a speaker vector).

- $M_{Z\to Y}$ is an image captioning (IC) that generates text captions (Y) from input images (Z),
- and $M_{Y\to Z}$ can be implemented as an image retrieval (IR) model that retrieves image (Z) given a text caption (Y) query or an image generation (IG) model that generates an image (Z) given a text caption (Y) input.

This chain implementation generalizes the chain mechanism when $|D| = 3$ by combining two chain implementations when $|D| = 2$. As illustrated in Figure 4(a), two loops are concatenated with text modality. The left-side loop is Tjandra *et al.*'s speech chain [2]–[5], which is connected with our proposed visual chain (IC and IR/IG) by text modality. We call this multimodal chain **MMC1-IR**, when the visual chain is using an IR model, and **MMC1-IG**, when the visual chain is using an IG model.

We designed the following training steps for MMC1-IR and MMC1-IG:

- **Step 1: supervised training with paired data**
  Each model is trained with a small amount of paired image-speech-text P_{xyz} data in a supervised manner. (Sec. II-A)
- **Step 2: semi-supervised training using unpaired data**
  The training can be continued in a semi-supervised manner using unpaired image-speech-text data $U_{x,y,z}$, as described in Sec. II-B. In the speech chain, $\mathcal{C}_{YXY}$ is the unsupervised step to train the $M_{X\to Y}$ ASR model using the reconstruction loss from the U_x data, and $\mathcal{C}_{XYX}$ is the unsupervised step to train the $M_{Y\to X}$ TTS model using the reconstruction loss from the U_Y data. In the visual chain, $\mathcal{C}_{ZYZ}$ and $\mathcal{C}_{YZY}$ are the unsupervised steps to improve the $M_{Y\to Z}$ IR/IG model and $M_{Z\to Y}$ IC models.
- **Step 3: semi-supervised training using single modality data**

Given speech only data S_x, we can make two chain paths: $\mathcal{C}_{XYX}$ and $\mathcal{C}_{XYZY}$. The first chain path ($\mathcal{C}_{XYX}$) can be used to train the TTS model using reconstruction loss $L_{M_{Y\to X}}$. In chain path $\mathcal{C}_{XYZY}$, the $\hat{y}$ transcription hypothesis is generated by the $M_{X\to Y}$ ASR model. Then this caption hypothesis is used by the $M_{Y\to Z}$ IR/IG model to produce image hypothesis $\hat{z}$, which can be used to generate a caption hypothesis $\hat{\hat{y}}$ by $M_{Z\to Y}$ IC model. $L_{M_{Z\to Y}}$ reconstruction loss can be calculated by comparing $\hat{y}$ and $\hat{\hat{y}}$, which then can be used to backpropagate the $M_{Z\to Y}$ IC model.

On the other hand, given image only data S_z, we can make two kinds of chain paths: $\mathcal{C}_{ZYZ}$ and $\mathcal{C}_{ZYXY}$. Path $\mathcal{C}_{ZYZ}$ trains the IR/IG model through the image's reconstruction loss. We emphasize path $\mathcal{C}_{ZYXY}$ that trains the $M_{X\to Y}$ ASR model, which generates transcription hypothesis $\hat{\hat{y}}$ that is transcribed from the $\hat{x}$ speech hypothesis generated by the $M_{Y\to X}$ TTS model. Then reconstruction loss $L_{M_{X\to Y}}$ can be calculated by comparing the transcription hypothesis $\hat{\hat{y}}$ with caption hypothesis $\hat{y}$ generated from the $M_{Z\to Y}$ IC model from image input z. Our main interest is determining whether the ASR model can be improved even with the image-only dataset, which has unrelated modality (text-speech) with ASR.

B. SINGLE-LOOP MULTIMODAL CHAIN (MMC2)

Next, we proposed a single-loop multimodal chain (MMC2) to show the implementation of our proposed chain framework in a multi-source multimodal model environment. In this multimodal chain, we combined ASR and IC to promote sharing between these two models. Therefore, the loop mechanism resembles a chain implementation when $|D| = 2$ (Section II-B), although it can still process data with three kinds of modalities ($|D| = 3$):

- $M_{\{X,Z\}\rightarrow Y}$ is implemented as the ImgSp2Txt model that transcribes speech or caption images when given speech (X), images (Z), or both (XZ),
- $M_{Y\rightarrow X}$ is a TTS model that synthesizes speech (X) from text (Y),
- and $M_{Y\rightarrow Z}$ is an IG model that generates an image (Z) given a text caption (Y) input.

As illustrated in Figure 4(b), there is only one loop as the result of introducing ImgSp2Txt. This ImgSp2Txt model can be trained with image-speech, image only, or speech only data.

- **Step 1: cross-modal model supervised training**
  When paired image-speech-text data P_{xyz} are available, ImgSp2Txt can be trained in supervised manner.
- **Steps 2 & 3: semi-supervised training using unpaired and single modality data**
  The MMC2 has a different semi-supervised step because it operates in a single-loop mechanism. To adapt it into the chain path notation, let us assume $G = \{X, Z, XZ\}$. Then the IG or TTS model is $M_{Y\rightarrow G}$, depending on the desired output. Therefore, we can define two chain paths: $\mathcal{C}_{GYG}$ and $\mathcal{C}_{YGY}$, resembling chain paths when $|D| = 2$.
  The first path $\mathcal{C}_{GYG}$ is used when MMC2 is given either unpaired image-speech-text dataset $U_{x,y,z}$, speech-only dataset S_x, or image-only dataset S_z. The ImgSp2Txt model generates text hypothesis $\hat{y}$ so that either IG or TTS can generate an image or speech depending on the input. If the input is an image, the IG can be back-propagated by the reconstruction loss from the image hypothesis generated by IG. When the input is speech-only, the ImgSp2Txt model generates text hypothesis $\hat{y}$, which is used by TTS to generate speech $\hat{x}$. By comparing the generated and original speech in the TTS reconstruction loss, we can backpropagate the TTS model. Then for the second chain path $\mathcal{C}_{YGY}$, both TTS and IG produce speech and image from the text input. These speech and images then can be used to backpropagate the $M_{G\rightarrow Y}$ ImgSp2Txt model using the reconstruction loss $L_{M_{G\rightarrow Y}}$ by comparing the original text y and text hypothesis $\hat{y}$.

V. MULTIMODAL CHAIN COMPONENTS

In this section we listed all the components of multimodal chain.

A. SEQUENCE-TO-SEQUENCE ASR

We build the sequence-to-sequence ASR model that resembles the Listen, Attend, and Spell (LAS) framework, which uses location-aware attention [12]. As illustrated in Figure 5, this model encodes a speech feature $\mathbf{x} = [x_0, \ldots, x_s]$ with bidirectional long-short term memory (LSTM) layers into a speech embedded representation $\mathbf{e^{ASR}} = [e_0^{ASR}, \ldots, e_s^{ASR}]$ which is a high-level feature representation used for decoder.

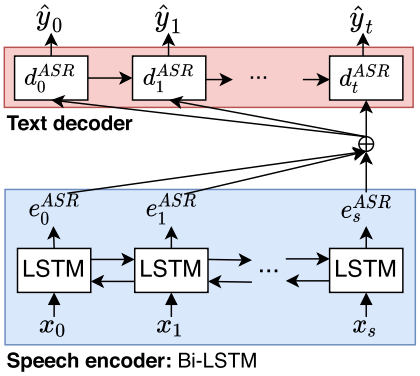


FIGURE 5. ASR model.

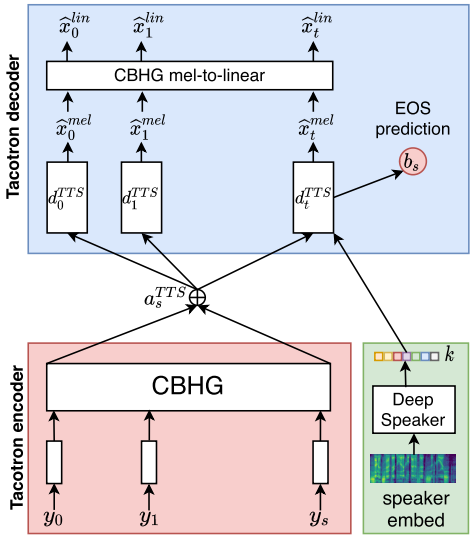


FIGURE 6. TTS model.

Then, the decoder uses teacher-forcing against text sequence $\mathbf{y} = [y_0, \ldots, y_t]$ with length t to generate a sequence of text hypothesis $\hat{\mathbf{y}} = [\hat{y}_0, \ldots, \hat{y}_t]$. To condition this generation process against speech features, we used an attention mechanism [13] to produce alignment probability $a_t = Align(e^{ASR}, d_t^{ASR})$, given encoded representation e^{ASR} and decoder hidden state d_t^{ASR}. Then the alignment probability are used to weight the encoded representation producing a context vector c_t, so that the hypothesis probability can be produced by an output layer $p_t = out([c_t, d_t^{ASR}])$.

B. SEQUENCE-TO-SEQUENCE TTS

A sequence-to-sequence TTS receives a text utterance $\mathbf{y} = [y_0, \ldots, y_s]$ and learn to generate a speech feature $\mathbf{x} = [x_0, \ldots, x_t]$ by optimizing its parameters. We build our model based on a sequence-to-sequence TTS with a one-shot speaker adaptation [3], which was adapted from the basic structure of the Tacotron TTS [14] and DeepSpeaker [15] models (Figure 6).

In this model, the speech feature generation in the decoder part is conditioned not only on text utterance $\mathbf{y}$ but also on speaker embedding k. These speaker embeddings are randomly sampled from the pool of speaker embeddings extracted from the paired speech-text data using a

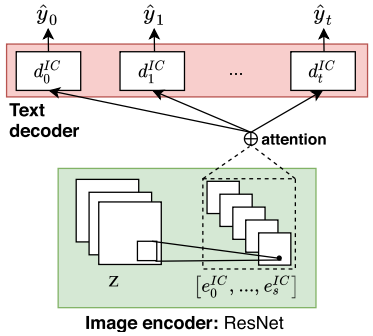


FIGURE 7. IC model.

DeepSpeaker model that was trained on the same data. We used an end-of-speech (EOS) prediction module [3] to predict the duration of the generated speech.

C. IMAGE CAPTIONING

An attention-based image captioning model encodes image $\mathbf{z}$ into high-level features $[e_0^{IC}, \ldots, e_s^{IC}]$ (Figure 7). These features are then used as the context for an attentional text decoder to generate hypothesis captions. To get this two-dimensional high-level features, a partial image classification model is usually used by taking the hidden representation before the pooling operation. Then these features are attended by a multilayer perceptron attention module which produces alignment probability $a_t = Align(e_s^{IC}, d_t^{IC})$ given encoded representation e_s^{IC} and decoder hidden state d_t^{IC}. Then the alignment probability is used to weight the encoded representation producing context vector c_t. By the hypothesis probability of each timestep $p_t = out([c_t, d_t^{IC}])$, the decoder then decodes a sequence of caption hypotheses using teacher-forcing against the original text sequence. In this study, we use ResNet [16] as image encoder, and LSTM decoder to decode text, resembling similar architecture with Xu *et al.* (2015) who proposed the "Show, Attend, and Tell" model [17].

D. IMAGE RETRIEVAL

An image retrieval model encodes image $\mathbf{z}$ and text caption $\mathbf{y}$ into embedding vectors v^z and v^y (Figure 8). The image encoder is usually constructed by a series of pretrained convolutional neural networks, followed by pooling and linear transformation at the end to produce image embedding v^z. Recurrent neural network is used to encode the text sequence into an embedding v^y. To combine both the image and text embeddings into a unique multimodal embedding space, we use a ranking loss with distance d that defines the distancing between positive (v^y, v^z) and negative samples $(\hat{v^y}, \hat{v^z})$. In this study, we use pairwise rank loss as the loss for image retrieval L_{IR} as follows:

$$L_{IR} = \sum_{|v^y|}\sum_{|\hat{v}^z|} max\{0, M + d(v^y, v^z) - d(v^y, \hat{v}^z)\} + \sum_{|v^z|}\sum_{|\hat{v}^y|} max\{0, M + d(v^z, v^y) - d(v^z, \hat{v}^y)\} \quad (14)$$

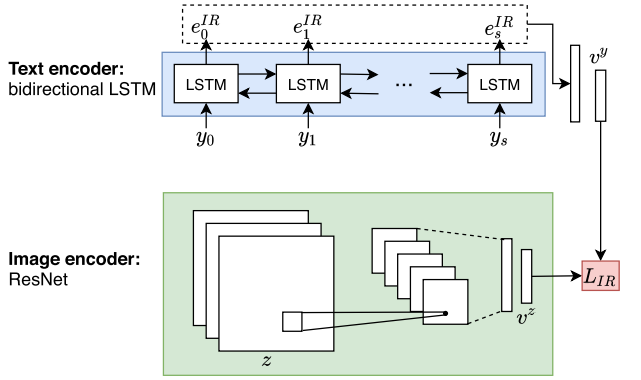


FIGURE 8. IR model.

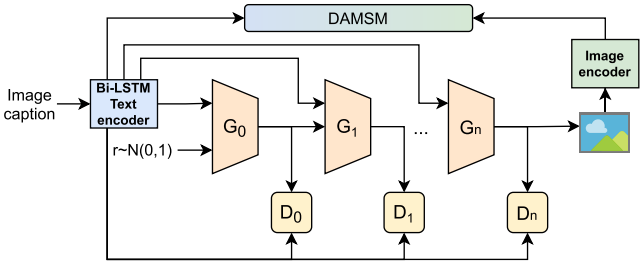


FIGURE 9. AttnGAN [18].

E. IMAGE GENERATION

AttnGAN [18] model generates images from a caption in a multistage manner, given a deep attentional multimodal similarity model (DAMSM). This multistage generating and discriminating strategy can successfully synthesize image in detailed clarity that are accurate to the given caption.

As illustrated in Figure 9, a bidirectional LSTM text encoder encodes the given image caption. Then its sentence vector is used as a condition to generate the image in the first stage using the G_0 generator model given the sentence vector and vector r that is sampled from a standard normal distribution. Then the generated image is evaluated using discriminator D_0. This process is repeated with DAMSM attention over the caption so that $[G_1, \ldots, G_n]$ generates images and $[D_1, \ldots, D_n]$ iteratively evaluates them until step n when the target image size has been reached.

F. TWOFOLD IMAGE-SPEECH TO TEXT MODEL

An image contains the information being spoken in its speech captions. We designed a single model that does both tasks to exploit this relation in the ASR and IC tasks. In addition, the model should be able to separately process speech and images if one of them is not available. When the input is only speech, this model will produce the transcription of the speech. An image caption is generated when only an image is provided. Finally, the model produces a speech transcription with the help of the input image when both image and speech are provided.

We designed output layer probability sharing between ASR and IC in a sequence-to-sequence ImgSp2Txt with a dual-decoder model (Figure 10). In this model, the image is encoded by a residual network that produces high-level

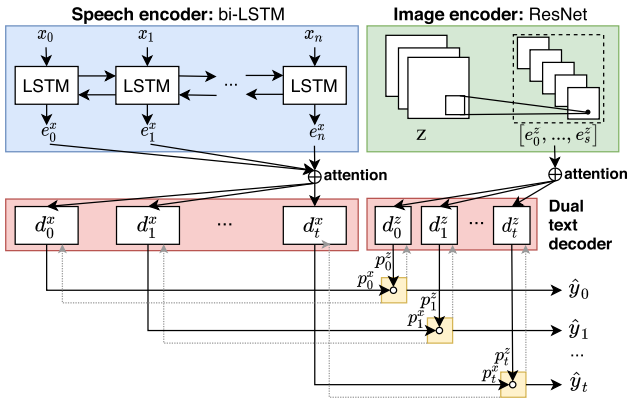


FIGURE 10. Dual text decoder with audio and visual decoding combination.

feature representation $\mathbf{e^z} = [e_0^z, \ldots, e_n^z]$ of the image. Bidirectional LSTM encodes the speech features into embedded representation $\mathbf{e^x} = [e_0^x, \ldots, e_m^x]$. Then the dual text decoder attends both $\mathbf{e^x}$ and $\mathbf{e^z}$. In training, softmax cross entropy loss $L_{ImgSp2Txt}$ is calculated by previously averaging both the p_t^y and p_t^z output layer probability for the image and speech input. If only one is available, the output layer probability of the respective modality is used.

VI. EXPERIMENT SET-UP

A. DATASET

1) Flickr 30k
Flickr30k [19] is an image-captioning dataset which images from Flickr consist of everyday activities, scenes, and events. There are about 150k crowd-sourced captions with 30k images in this dataset, which makes every image has 5 captions. Since this dataset only has text as a caption, we generated speech captions based on them using the Google TTS.

2) Flickr 8k
Similar to Flickr30k, Flickr8k [20] contains 8k images from Flickr. Each image has five captions annotated using the crowd-sourcing method. In addition, to enable the use of this corpus in the speech processing field, it was extended with natural speech recording using the Amazon Mechanical Turk crowd-sourcing platform. This dataset has 183 unique speakers.

B. DATASET COMPOSITION

We used the default dataset split for Flickr30k (29k train, 1k dev, and 1k test) and Flickr8k (6k train, 1k dev, and 1k test). However, because we proposed a semi-supervised strategy where a model quality can still be improved with just a single modality dataset, we designed a scenario that enables it. For Flickr8k, we used all the five captions from an image, while in Flickr30k, we follow the same settings as the previous work [10] to balance the image production side.

Table 1 lists each possible data modality type that we used in this study. Each modality type corresponds to a different training step depending on the scenario to be examined. The first type is all-paired modality type P_{xyz}, which contains

TABLE 1. Modality type with three conditions: (1) available paired data denoted as ◯, (2) available but unpaired data denoted as ▲, and unavailable data denoted as ×.

Modality type	sp x	txt y	img z	Description
P_{xyz}	◯	◯	◯	Multimodal paired
$U_{x,y,z}$	▲	▲	▲	Multimodal unpaired
S_x	▲	×	×	Single modality data (Speech only)
S_z	×	×	▲	Single modality data (Image only)

triplets of speech, text, and image. This type of data typically has the lowest number of data compared with other types in the dataset, because in this study we want to minimize the need for paired data as much as possible. Modality type $U_{x,y,z}$ means that all three modalities (speech, image and text) are available, but they are unpaired. Finally, modality type S_x and S_y are single-modality data that contain only speech and image modality respectively.

We partition the data into several subsets based on the modality type. Depending on the task, the number of data in each subset is different, as shown in Table 2. Before partitioning the data, we randomly shuffle the order of the keys in the dataset initially. For measuring the topline performance such as in Sec. VII-A, we assume that all data are paired. In this way, we can compare each of our model performance in supervised mode with other previously published studies.

To prove our hypothesis that improving ASR with image data is possible with a multimodal chain, we composed the following data partition on Flickr8k and Flickr30k. Paired data P_{xyz} has the smallest amount of data, followed by unpaired data $U_{x,y,z}$ and S_x, S_z, which comprises the largest portion. First, we trained the ASR, TTS, IC, IR, and IG models with this data partition, following Steps 1-3 from Section IV. With these trained models, we can compare which image production method is better for the multimodal chain: IR or IG. As listed in Table 2, we use the Flickr30k dataset with 2000 P_{xyz} data, 7000 $U_{x,y,z}$ data, and 10000 S_x, S_z data.

We used Flickr8k with 800 P_{xyz} data, 1500 $U_{x,y,z}$ data, and 1850 S_x, S_z data to show that our proposed multimodal chain can also work in a multi-speaker natural speech dataset. We tested MMC2 with the same data partition to compare it with a label propagation method (Section VI-D). We also tested what happens when all the remaining data (other than the paired P_{xyz} data) are unpaired or single modality.

We designed the data partition to verify the effect of the amount of single modality data on the final performance. Using a model supervisedly trained with 800 P_{xyz} data, we continued the training in a semi-supervised manner based on Step 3 (Section II-C). The remaining data (other than the paired data) were regarded as single modality, which we divided into 2600 S_x speech-only data and 2600 S_z image-only data. We ran the experiment with a variable amount of single modality data to identify the correlation between the data amount and the final speech processing model performance.

Finally, to see the initial data amount effect of the final speech processing model's improvement, we variably changed the amount of paired P_{xyz} data. After that,

TABLE 2. Data partitioning for each subset (in #Image). $n = \{0, 1, 2, 3, 4, 5\}$, $m = \{0, 1, 2, 3, 4, 5, 6, 7\}$.

Task	Note	Section	Paired $\#P_{xyz}$	Unpaired $\#U_{x,y,z}$	Single-modality $\#S_x$	Single-modality $\#S_z$	Total
Dataset: Flickr30k							
Topline	For comparison with existing published systems	VII-A	29000	0	0	0	29000
IR vs IG	For comparison between MMC1-IR and MMC1-IG	VII-B	2000	7000	10000	10000	29000
Dataset: Flickr8k							
Topline	For comparison with existing published systems	VII-A	6000	0	0	0	6000
(Label Prop. I) & (MMC1 vs MMC2)	For comparison between label propagation and our proposed multimodal chain	VII-C&VII-D	800	1500	1850	1850	6000
Label Prop. II	For comparison with label propagation using more paired data	VII-C	1400	900	1850	1850	6000
No single modality	For checking performances when all remaining data other than paired are unpaired	VII-D	800	5200	0	0	6000
Var. single modality	For investigation of effect of increasing single-modality data amount	VII-E	800	0	$520n$	$520n$	800-6000
Var. paired	For investigating the effect of increasing initial paired data amount	VII-F	$200+300m$	0	1850	1850	3900-6000

we continued the training using a fixed amount of 1850 S_x and 1850 S_z single modality data. The interesting point here is how much the initial model performance improved.

The image-only dataset cannot be used in all scenarios without our proposed multimodal chain training strategy, which implies that no further improvement to the existing speech chain can be done. Therefore, our main interest here is to determine whether ASR improvement remains possible even when only image data are available.

C. MODEL DETAILS AND EVALUATION METRICS

We used all the models described in Sec. V. We used an 80-dimensional mel-spectrogram for the speech features in the ASR and TTS models. We used word-level text granularity for the IC and IG tasks, and the ASR, TTS, and ImgSp2Txt used character-level granularity. To handle the unseen speakers in the unpaired and single modality data, we performed one-shot speaker adaptation [3] with a modified speaker embedding size from 128 to 64. We decoded all the hypotheses during the chain connection and in testing using a beam size of three. We used Adam optimizer for all the models except IR; with 1e-3 for the ASR, 1e-4 for the ImgSp2Txt, and IC models, 2.5e-4 for the TTS model, and 2e-4 for the IG models. For the IR model, we used a stochastic gradient descent with a 0.1 learning rate. Although technically training each element in the chain path is possible with an end-to-end style [4], we discovered that the gradient for the early components of the chain became too small for a long chain path. Therefore, in this study, we just backpropagated the last model of the chain path.

For the ASR and TTS models, we used a similar model parameter as Tjandra *et al.* (2017) [3]. For IC model, we removed the last two layers of the ResNet [16], which yielded a grid of high-level representation to each image region it represents. We used an LSTM decoder with a 512 hidden size to decode the hypothesis captions. For the IR model, we removed the last layer of ResNet, a step that yielded a vector of the image itself. These representations were linearly transformed to 300-dimensional image embeddings. We generated sentence embeddings with a bidirectional LSTM with 256 hidden sizes in each direction. We used the same parameters as Xu *et al.* (2018) for the AttnGAN. In practice, to reduce the memory usage, IC, IR, and IG models used only a 128×128 pixels image size.

We evaluated each model with the test set of the dataset with which it was trained. We measured the ASR performance with the character error rate or the word error rate (CER/WER) and a bilingual evaluation understudy (BLEU) [21] for the IC to compare the n-gram between the hypothesis and the reference captions. We used 1-gram and 4-grams for BLEU, denoted as B1 and B4. In addition, to measure the TTS performance, we used L2-norm^2 metrics (denoted as L2) to measure the error between the reference and generated mel-spectrogram sequences. Finally, IG was measured by inception score (IS) [22] to determine how realistic the IG output was.

D. LABEL PROPAGATION

Label propagation [23] is a common semi-supervised training strategy that generates pseudo labels from partially unlabeled data. In its deep neural network implementation, this kind of approach is also known as pseudo labels [24]. We adapted this algorithm to follow our use cases. First, a model is trained with the labeled portion of the data. Then the trained model generates a pseudo label for the unlabeled data. To use this as a baseline in our task, we modified the algorithm to use it for the cross-modal tasks.

Assume that an ASR model is trained using the speech and text parts of the P_x data subset. Then the text part of the $U_{x,y,z}$ subset is generated using the trained model. These text hypotheses are used to retrain the model. This process is repeated for the data in the U_x type subset. The same process can be used for the IC model with the S_z type subset. However, for IG and TTS, the last step using the S_x and S_z type subsets cannot be used because the data modality is on the target side. To solve this problem, we generate source side data using the corresponding model. For example, to use the speech only S_x type subset for TTS, we generate a text hypothesis using the ASR model on that same step.

VII. EXPERIMENT RESULTS

A. TOPLINE SCENARIO

In this section, we simulated a condition where much paired data are available (Table 2: Topline). Our experiment

TABLE 3. Comparison of our model performances with existing published results: ↓ means lower is better; ↑ means higher is better.

Data	Model	Result
ASR - CER (%) ↓		
WSJ [25]	Kim et al. [26]	11.08
	Tjandra et al. [4], [27]	6.43
	Ours (Sec. V-A)	6.60
TTS - L2 ↓		
WSJ	Tjandra et al. [27]	0.64
	Ours (Sec. V-B)	0.68
IC - B1/B4 ↑		
Flickr8k	Xu et al. [17]	66.90 / 19.90
	Ours (Sec. V-C)	65.93 / 22.56
IR - R@10 ↑		
Flickr30k	Vilalta et al. [28]	59.8
	Ours (Sec. V-D)	62.42
IG - Inception ↑		
CUB	Xu et al. [18]	4.36
	Ours (Sec. V-E)	5.67
ImgSp2Txt - CER / WER (%) ↓		
Flickr8k	Sun et al. [29]	- / 13.81
	Ours (Sec. V-F)	5.16 / 7.13

measured our model performance and compared it to previously published studies. When no previously published result was available for the Flickr8k or Flickr30k datasets, we trained our model with the same dataset that was used in the reported result. Therefore, since each model result reported in each section (i.e., ASR, TTS) was trained with the same dataset, they are comparable.

We listed all the scores of our topline model in Table 3. In the WSJ corpus [25], our ASR model outperformed better than Kim *et al.* [26], and both ASR and TTS work as well as the previously published results of Tjandra *et al.* [4], [27]. Our IC model performed better than Xu *et al.* [17] in BLEU4. We also observed similar performance in our IR model in the Flickr30k dataset, the IG model in the CUB dataset, and the ImgSp2Txt model in Flickr8k.

B. PROPOSED: FROM IR TO IG

First, we need to decide whether the IR or the IG model is better for the multimodal chain. The benefit of using IR is that the retrieved image is of good quality because no synthesis is needed. However, because the image is retrieved, it is difficult to return unseen images, especially when the dataset is not parallel. On the other hand, generating images using the IG model produced better unseen images because they are synthesized. Even so, the image quality is not ideal, especially for the open-domain dataset in this study.

For this, we used MMC1 and replaced the image production model using IR or IG. We labeled each of them as MMC1-IR and MMC1-IG. We partitioned the data for Steps 1, 2, and 3 following the steps in Section IV. For the size of each subset, refer to Table 2: ''IR vs IG''. We trained all initial model in a supervised manner with paired P_{xyz} type data subset and semi-supervisedly trained the model inside the speech and visual chains using $U_{x,y,z}$ type subset. As shown in Table 4, although both the ASR and TTS models are unaffected because there is no influence from the image production model yet, the IC performance between IR and IG in the visual chain can already be compared. The MMC1-IR improvement in Step 2 is more focused on B1 than B4, compared with MMC1-IG, which consistently improves both. For the image production models, both IR and IG show improvement in their own evaluation metrics.

Next, we connected the speech and visual chains using text modality in Step 3. All the speech processing models in MMC1-IG outperformed MMC1-IR, showing that a visual chain using IG can generate a better text hypothesis to be fed into a speech chain than with IR. This result can be quantitatively compared in the IC score, where MMC1-IR shows a performance decrease, although in MMC1-IG both the B1 and B4 scores increased. We also observed a decrease in the IR model performance. In this step, the IR model receives text hypotheses generated by the ASR model from the S_x speech-only data subset. Unfortunately, when the IR model needs to retrieve images for these text hypotheses, it can only get images from the U_z and S_z type data subsets. These data don't have exact matches for such transcribed S_x type data (S_x and S_z are not parallel), which lead us to infer that the MMC1-IR is struggling to retrieve unseen images. Although it is possible to use Hybrid IR + IG (i.e., IR for Step 2 and IG for Step 3), we decided that this step is inefficient because we need to train both the IR and IG models. Due to these considerations, we decided to use the IG model for our next experiments.

C. BASELINE: LABEL PROPAGATION

In this section, we did label propagation to learn how much improvement we can get with identical data composition. We call this experiment Label Propagation I, whose results are shown in Table 5. By using the same amount of initial data, the ASR, IC, and ImgSp2txt models cannot be improved, although some improvement was reported in the TTS and IG task.

To investigate whether more data can raise the improvement, we added more paired data to the initial step by taking 600 images from the unpaired multimodal data in Step 2 and called this experiment Label Propagation II. By using this new composition, the ASR performance can be maintained, and we found improvement in the other models. Compared with our proposed multimodal chain, even with less paired data, such as in Label Propagation I, all of the models can still be improved. This result shows that our proposed multimodal chain is more effective than the label propagation method.

D. PROPOSED: COMPARING MMC1-IG AND MMC2

After choosing between IR and IG and comparing with the label propagation baseline, in this section we evaluate the performance of a dual-loop (MMC1-IG) vs. a single-loop multimodal chain (MMC2). For the data partitioning in this experiment, we refer to the subset partitioning based in Table 2: MMC1 vs MMC2. Initially, we separately trained all the models using P_{xyz} data in a supervised manner. As shown in Table 5, both MMC1-IG and MMC2 have

TABLE 4. Comparison of performance of proposed MMC1-IR with MMC1-IG on Flickr30k.

Training	Data Type	#Image	ASR	IC	TTS	IR	IG
			CER↓	B1/B4↑	$L2^2$↓	R@10↑	IS↑
MMC1-IR	P_{xyz} Multimodal	2000	21.46	45.97/10.55	0.72	14.30	-
	$+U_{x,y,z}$ Multimodal	7000	4.02	48.00/10.08	0.49	16.08	-
	$+S_{x,z}$ Sp/Img only	10000	3.51	47.60/9.82	0.44	15.50	-
MMC1-IG	P_{xyz} Multimodal	2000	21.46	45.97/10.55	0.72	-	4.06
	$+U_{x,y,z}$ Multimodal	7000	4.02	46.55/10.92	0.49	-	5.59
	$+S_{x,z}$ Sp/Img only	10000	2.77	47.33/11.38	0.43	-	7.21
Topline	P_{xyz} Multimodal	29000	0.68	51.34/13.64	0.40	40.22	7.57

TABLE 5. Comparison of proposed MMC1 and MMC2 performances with label propagation method in Flickr8k dataset.

Training	Data Type	#Image	MMC1-IG				MMC2			
			ASR	IC	TTS	IG	ImgSp2Txt		TTS	IG
			CER↓	B4↑	$L2^2$↓	IS↑	CER↓	B4↑	$L2^2$↓	IS↑
Label Propagation I	P_{xyz} Multimodal	800	36.35	12.75	0.77	5.90	26.67	32.23	0.77	5.90
(Semi-Supervised)	$+U_{x,y,z}$ Multimodal	1500	39.57	12.53	0.77	7.04	27.45	33.59	0.77	7.04
	$+S_x$ Sp only	1850	46.04	-	0.63	-	28.87	35.75	0.63	-
	$+S_z$ Img only	1850	-	11.41	-	7.20	30.31	35.38	-	7.20
Label Propagation II	P_{xyz} Multimodal	800+α	15.52	15.10	0.64	7.25	13.54	57.63	0.64	7.25
Plus $\alpha = 600$	$+U_{x,y,z}$ Multimodal	1500-α	15.36	15.63	0.62	7.82	13.22	58.66	0.62	7.82
(Semi-Supervised)	$+S_x$ Sp only	1850	15.28	-	0.55	-	14.36	59.36	0.55	-
	$+S_z$ Img only	1850	-	15.86	-	8.86	15.24	58.69	-	8.86
Proposed	P_{xyz} Multimodal	800	36.35	12.75	0.77	5.90	26.67	32.23	0.77	5.90
Multimodal Chain	$+U_{x,y,z}$ Multimodal	1500	15.10	13.22	0.59	8.29	14.88	55.15	0.65	10.12
(Semi-Supervised)	$+S_z$ Img only	1850	12.70	14.11	0.60	9.58	13.74	58.65	0.64	10.00
img → sp	$+S_x$ Sp only	1850	12.39	13.88	0.56	9.03	12.84	59.61	0.62	10.40
Proposed	P_{xyz} Multimodal	800	36.35	12.75	0.77	5.90	26.67	32.23	0.77	5.90
Multimodal Chain	$+U_{x,y,z}$ Multimodal	1500	15.10	13.22	0.59	8.29	14.88	55.15	0.65	10.12
(Semi-Supervised)	$+S_x$ Sp only	1850	12.37	13.28	0.56	9.12	13.81	58.03	0.62	10.65
sp → img	$+S_z$ Img only	1850	12.06	13.29	0.56	9.11	12.32	59.66	0.61	9.95
Proposed	P_{xyz} Multimodal	800	36.35	12.75	0.77	5.90	26.67	32.23	0.77	5.90
Separated (Semi-supervised)	$+U_{x,y,z}$ Multimodal	5200	10.48	14.23	0.53	6.29	13.88	58.60	0.63	9.45
Topline (Supervised)	P_{xyz} Multimodal	6000	5.76	19.91	0.50	9.66	5.16	79.88	0.50	9.66

identical TTS and IG scores because they are using the same initial model. ImgSp2Txt has a better CER score than ASR for this initial step because ImgSp2Txt combines image and speech information using a multi-source model.

We continued the training of these initial models using the $U_{x,y,z}$ data subset, and both MMC1-IG and MMC2 showed improvement for all models. We separated the use of data based on the modality of Step 3 to understand how specific modality contributes to the improvement of each chain component. First, we started training with image-only data S_z and continued with speech-only data S_x ($img \rightarrow sp$). For comparison, we also trained with speech-only data S_x first and continued with image-only data S_z ($sp \rightarrow img$). As shown in Table 5, in terms of ASR performance, the $sp \rightarrow img$ combination is more effective. By training with image-only data, we observed improvement not only in the image-processing related task but also in the speech processing model. This shows that the cross-modal augmentation inside the chain is effective, either in a dual-loop MMC1 or in a single-loop MMC2.

Next, we measured the actual effectiveness of the cross-modal augmentation inside the chain by separately training each speech and visual chain. We assume that except for the 800 paired data P_{xyz}, all the other 5200 data ($U_{x,y,z}$) are unpaired. Therefore, each chain gets a hypothesis from its related modalities, unlike our proposed multimodal chain. For MMC1-IG, this approach yield 10.48% CER which is 1.58 points better than the best approach of 12.06% when some data have only single modality (See Table 5: Separated(Semi-supervised)). In a single-loop MMC2, however, our proposed method remains superior. With our proposed multimodal chain, we can improve the ASR performance with unrelated modality data (image) to a decent level through cross-modal augmentation, even when the speech and image datasets are disjointed.

We also listed the result when we assumed that all the data are paired. This result shows the distance between our proposed semi-supervised approach and the supervised approach. Finally, we compared our best semi-supervised ASR performance (12.06% CER/17.84% WER), which is comparable to Sun *et al.*'s supervised ASR, which has a 13.81% WER on the same Flickr8k dataset [29]. Although our proposed approach is semi-supervised, we can still achieve a comparable error rate to a fully-supervised ASR system.

E. SINGLE MODALITY DATA AMOUNT EFFECT TO THE FINAL SPEECH PROCESSING MODEL PERFORMANCE

Our proposed multimodal chain emphasizes its ability to produce additional improvement in speech processing models even when no more speech or text data are available. Therefore, we investigated whether speech processing models

TABLE 6. ASR performance improvement given various initial data amount in Flickr8k natural speech dataset.

P_{xyz} 200+m300 images	$+S_{xz}$ sp/img only	P_{xyz} CER *initial model*	$+S_{x,z}$ CER *multimodal chain*	△CER ↑ *multimodal chain*
6000 (all training subsets)	0	5.76	n/a	n/a
2300 (m=7)	1850	9.97	10.05	-0.08
2000 (m=6)	1850	11.54	11.54	0.00
1700 (m=5)	1850	13.42	13.02	0.40
1400 (m=4)	1850	15.52	14.62	0.90
1100 (m=3)	1850	19.13	18.00	1.13
800 (m=2)	1850	36.35	25.35	11.00
500 (m=1)	1850	77.65	48.41	29.24
200 (m=0)	1850	77.45	72.93	4.52

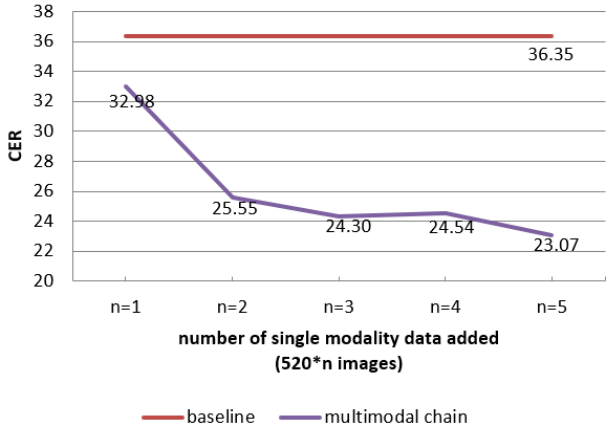


FIGURE 11. Single modality data amount effect to final ASR performance compared with initial model baseline in Flickr8k natural speech dataset. Vertical axis: character error rate (CER). Horizontal axis: number of single modality data added.

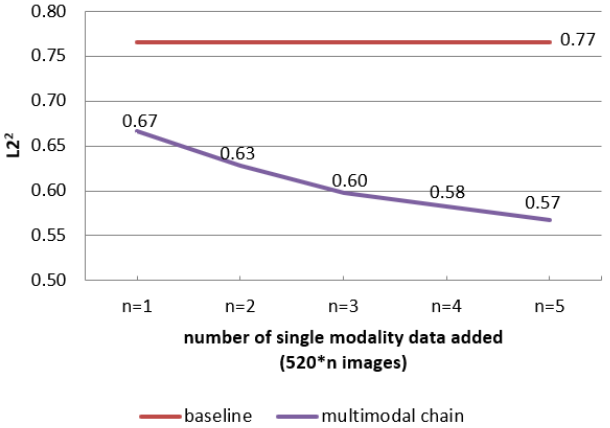


FIGURE 12. Single modality data amount effect to final TTS performance compared with initial model baseline in Flickr8k natural speech dataset. Vertical axis: L2^2 Loss. Horizontal axis: number of single modality data added.

improve consistently as the amount of single modality data increases. For this additional experiment, we refer to the data partitioning shown in Table 2: Var. Single Modality.

Figure 11 compares the ASR improvement using multimodal chains with the initial model performance in terms of CER. The horizontal axis shows the number of single modality data types ($S_{x,z}$) added in 520-image increments. These increments generated five trained models, whose performances relatively decrease, given more data to the multimodal chain. The best CER score (23.07%) was reached using all of the single modality data of 2600 images.

In addition, Figure 12 compares the TTS improvement using multimodal chains with the label propagation method and the initial model performance in terms of L2^2 loss. Compared with ASR, the TTS performance is consistently better, given more single modality data. The best TTS performance was reached with the most single modality data of 2600 images, which yields 0.20 L2^2 loss improvement compared with the initial baseline. These results suggest that the improvement from multimodal chains is positively related to how many more data are used in the semi-supervised step by leveraging the cross-modal augmentation.

F. INITIAL DATA AMOUNT EFFECT TO FINAL SPEECH PROCESSING MODEL PERFORMANCE

In this section, we experimentally changed the amount of initial data used to supervisedly train the initial model with the data partitioning shown in Table 2: Var. Paired. We used data subset P_{xyz} variably to test the training with various initial data amounts. We continued the training with single modality data $S_{x,z}$. To measure the effectiveness of multimodal chains to improve the performance in semi-supervised steps, we measured the score differences between the initial model and the model after the multimodal chain semi-supervised step.

The ASR performance improvement can be seen in Table 6. Using all the training sets as initial data (6000 images), we got a 5.76% CER for the ASR performance. We reduced the amount of initial data and reserved the remaining data for the multimodal chain semi-supervised step. In this scenario, a larger amount of initial data denotes a better performance in the initial model. We used the same number of speech and image-only data for all the possible initial data. When the number of initial data was reduced to 1700 images, our proposed multimodal chain started to show its effectiveness in improving the ASR model performance, manifested by positive △CER scores. The highest performance increases were achieved with the initial data pair of 500 images.

However, that is not the case with the TTS model performance improvement (Table 7). The △L2^2 score remained positive when the initial data sizes exceed 500, suggesting that our proposed multimodal chain improved the TTS performance even when the initial TTS model was already

TABLE 7. TTS performance improvement given various initial data amount in Flickr8k natural speech dataset.

P_{xyz} 200+m300 images	$+S_{xz}$ sp/img only	P_{xyz} L2^2 *initial model*	$+S_{x,z}$ L2^2 *multimodal chain*	$\triangle$L2^2 $\uparrow$ *multimodal chain*
6000 (all training subsets)	0	0.50	n/a	n/a
2300 (m=7)	1850	0.57	0.55	0.02
2000 (m=6)	1850	0.70	0.57	0.12
1700 (m=5)	1850	0.61	0.55	0.05
1400 (m=4)	1850	0.64	0.56	0.08
1100 (m=3)	1850	0.66	0.59	0.07
800 (m=2)	1850	0.77	0.61	0.16
500 (m=1)	1850	0.78	0.71	0.06
200 (m=0)	1850	0.86	0.87	-0.01

relatively good. The experiment with an initial data amount of 200 images showed no improvement in terms of $\triangle$L2^2. Since the performance of the initial ASR and TTS models is too low, they cannot effectively assist each other inside the chain.

We also investigated the feasibility of using the existing pretrained model with the ASR and TTS model previously trained in the WSJ-SI284 dataset [25]. We continued the training of this pretrained model in a semi-supervised manner with the $S_{x,z}$ dataset in the same manner with the experiment in this section. We then tested it with the Flickr8k test set and found an 0.5% CER improvement from the 90.72% CER scores for the initial model. We conclude that although improvement exists, the domain similarity between the initial and single-modality datasets must be considered. The WSJ dataset consists of news domain utterances, and Flickr8k is an image caption dataset that contains declarative caption sentences that describe what is happening in the images. Therefore, these two datasets have very few contents overlaps.

From these experiments, we conclude that the accuracy of the initial model, which was trained in the first step, affects the final semi-supervised chain performance. We also found that our proposed multimodal chain is more effective in a low-data condition when the initial model can still provide a meaningful hypothesis to assist each other in the semi-supervised chain training process. Finally, focusing on ASR performance, we found that the initial paired data amount of 500 images gave the most improvement, and the one with 800 images gave a relatively better final CER.

VIII. RELATED WORKS

Many studies have integrated audio and visual information to improve speech recognition performance, including deep learning approaches. The first end-to-end approach for audiovisual speech recognition was proposed by Petridis *et al.* [30]. A popular extension of the LAS framework [12], called "Watch, Listen, Attend, and Spell (WLAS)," was proposed by Chung *et al.* [31]. This framework introduced a dual-attention mechanism to enable the processing of speech and/or images together depending on the data availability. Afouras *et al.* [32] also proposed a deep audio-visual speech recognition system to recognize phrases and sentences from a talking face. Recently, Wu *et al.* proposed a Dual Attention Matching (DAM) for audio-visual event localization [33]. However, most of these approaches are used in conditions where the video or face data are highly parallel to the speech or audio data, a context that creates a monotonic alignment between the visual and speech modalities.

Sun *et al.* [29] proposed a "Look, Listen, and Decode" model that uses photos to improve the ASR process in the Flickr8k dataset. This task is more challenging than lip-reading tasks because the audiovisual model needs to decide which part of the image is useful for the transcription task. However, by adding more modalities, such as images, collecting a dataset for this supervised task is complicated because a parallel triplet is needed: speech, text, and image.

Although adding more modality creates a more robust and flexible system, all these approaches need parallel data for supervised training. Herein lurks the difficulty; if a model is translating from one modality to another, it needs a paired tuplet of data so that it can be trained in a supervised manner. If we add another modality to the process, then we need a triplet of data, and so on. This phenomenon contributes to the difficulty of building a multimodal system.

To alleviate this limited parallel data problem by enabling training from singleton data, some methods have been proposed under the name of dual learning or cycle consistency. Dual learning in machine translation was proposed using the back-translation hypothesis of the dual model to generate pseudo labels to train a primal model [34]. In the speech-to-speech domain, a study converted synthetic to natural speech by cycle consistency [35]. Extensive studies are also available in the image to image domain, such as DiscoGAN [36], CycleGAN [37], and DualGAN [38].

The speech chain framework [2]–[5] might be the first framework constructed on different modality domains (speech versus text). In the image to text domain, Turbo Learning combined image captioning and generation in a joint training framework [39]. Recently, a multimodal machine chain [10] accommodated triangle modality and the loop feedback mechanism.

Research interest has been growing in the field of self-supervised learning. A self-supervised system solves a task by generating some kind of supervisory signals by itself. From the data perspective, the training data for self-supervised learning can be either automatically or approximately labeled. Several studies incorporated a self-supervised

approach with a semi-supervised learning task. Zhai *et al.* (2019) described how it is beneficial to introduce self-supervised loss in a semi-supervised task [40]. Si *et al.* (2020) proposed adversarial self-supervised learning for semi-supervised 3D action recognition [41]. On the other hand, Chen *et al.* (2020) showed that self-supervised models trained on a large amount of data can be further adapted for semi-supervised tasks. In this study, we regard reconstruction loss as a kind of self-supervision, where we can use the unlabeled data as they are. However, to reach that stage, the models inside our proposed chain must be weakly supervised beforehand using a small amount of labeled data. For these reasons, we classify our multimodal chain approach as a semi-supervised learning strategy with some self-supervision from leveraging cycle consistency.

IX. CONCLUSION AND FUTURE WORK

In this study, we defined a general framework for the universal chain problem. We developed a cross-modal model collaboration in the form of a closely-knitted chain that enables the use of unrelated modality data through weak supervision. We investigated the use of an adversarial image generation model to enable the generation of unseen images during the chain process. To enable multispeaker speech processing, we also implemented one-shot speaker adaptation. Then, we trained and tested our multimodal chain in a multispeaker natural speech dataset. Our chain mechanism can be implemented on an audiovisual model through a single-loop multimodal chain, without any significant performance decrease.

Our proposed approach outperforms the label propagation method. Speech processing components can be improved even when using the image-only dataset, which is enabled by our proposed multimodal chain mechanism. We also ran an experiment that determined the effectiveness of our proposed approach in accordance with the amount of data in the initial and semi-supervised steps. We found that our proposed multimodal chain is more effective in a low-resource scenario, when the initial paired data are insufficient to satisfiably train the cross-modal model.

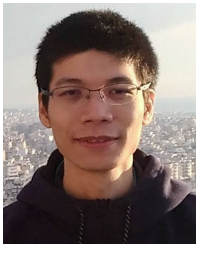
For future work, we will investigate the possibility of information sharing between each chain components to further reduce the amount of required paired data. We also want to develop better filtering or quality estimation during the passing of the hypothesis between the model inside the chain to improve the self-supervision within the chain. We are also interested in investigating domain adaptation strategies to enable cross-modal augmentation with data from different domains (i.e., news, travel, etc).

REFERENCES

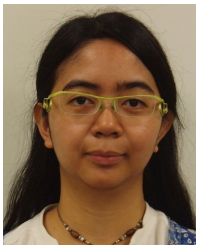
[1] P. Denes and E. Pinson, *The Speech Chain*. New York, NY, USA: Worth Publishers, 1993. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9020132
[2] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *Proc. ASRU*, 2017, pp. 301–308.
[3] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain with one-shot speaker adaptation," in *Proc. Interspeech*, Sep. 2018, pp. 887–891.
[4] A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end feedback loss in speech chain framework via straight-through estimator," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6281–6285.
[5] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 976–989, 2020.
[6] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, Mar. 1954.
[7] G. A. Calvert, "Crossmodal processing in the human brain: Insights from functional neuroimaging studies," *Cerebral Cortex*, vol. 11, no. 12, pp. 1110–1123, Dec. 2001, doi: 10.1093/cercor/11.12.1110.
[8] A. Byrne, "Recollection, perception, imagination," *Phil. Stud.*, vol. 148, no. 1, pp. 15–26, Mar. 2010.
[9] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. ICCV*, 2017, pp. 2070–2079.
[10] J. Effendi, A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking and visualizing: Improving ASR through multimodal chain," in *Proc. ASRU*, 2019, pp. 471–478.
[11] J. Effendi, A. Tjandra, S. Sakti, and S. Nakamura, "Augmenting images for ASR and TTS through single-loop and dual-loop multimodal chain framework," in *Proc. INTERSPEECH*, 2020, pp. 4901–4905.
[12] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.
[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473
[14] 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. INTERSPEECH*, 2017, pp. 4006–4010.
[15] 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," 2017, *arXiv:1705.02304*. [Online]. Available: http://arxiv.org/abs/1705.02304
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[17] X. Formatted, 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. ICML*, 2015, pp. 2048–2057.
[18] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.
[19] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr 30 k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.
[20] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. NAACL*, 2010, pp. 139–147.
[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.
[22] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, 2016, pp. 2234–2242.
[23] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," CMU, Pittsburgh, PA, USA, Tech. Rep. 107, 2002.
[24] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *Proc. ICML Workshop Challenges Represent. Learn. (WREPL)*, Jul. 2013, pp. 1–5. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.664.3543
[25] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop Held Harriman*, New York, NY, USA, Feb. 1992, pp. 1–2. [Online]. Available: https://www.aclweb.org/anthology/H92-1073
[26] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, 2017, pp. 4835–4839.

[27] A. Tjandra, S. Sakti, and S. Nakamura, "Multi-scale alignment and contextual history for attention mechanism in sequence-to-sequence model," in *Proc. SLT*, 2018, pp. 648–655.
[28] 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*, vol. 10, no. 5, pp. 900–923, 2019.
[29] F. Sun, D. Harwath, and J. Glass, "Look, listen, and decode: Multimodal speech recognition with images," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2016, pp. 573–578.
[30] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-End audiovisual fusion with LSTMs," in *Proc. 14th Int. Conf. Auditory-Visual Speech Process.*, Aug. 2017, pp. 36–40.
[31] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3444–3453.
[32] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2020, doi: 10.1109/TPAMI.2018.2889052. [Online]. Available: https://ieeexplore.ieee.org/document/8585066
[33] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6292–6300.
[34] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Proc. NIPS*, 2016, pp. 820–828.
[35] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, "Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 632–639.
[36] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.
[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[38] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for Image-to-Image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2868–2876.
[39] Q. Huang, P. Zhang, D. Wu, and L. Zhang, "Turbo learning for captionbot and drawingbot," 2018, *arXiv:1805.08170*. [Online]. Available: http://arxiv.org/abs/1805.08170
[40] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4L: Self-supervised semi-supervised learning," 2019, *arXiv:1905.03670*. [Online]. Available: http://arxiv.org/abs/1905.03670
[41] C. Si, X. Nie, W. Wang, L. Wang, T. Tan, and J. Feng, "Adversarial self-supervised learning for semi-supervised 3D action recognition," in *Proc. ECCV*, 2020, pp. 35–51.

JOHANES EFFENDI received the B.S. degree *(cum laude)* in computer science from Universitas Indonesia, Indonesia, in 2015, and the M.Eng. degree from the Nara Institute of Science and Technology (NAIST), Japan, in 2018, where he is currently pursuing the Ph.D. degree with the Augmented Human Communication Laboratory. His research interests include speech processing, neural machine translation, and natural language processing. He is a recipient of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Scholarship.



ANDROS TJANDRA received the B.Sc. and M.Sc. degrees *(cum laude)* from the Faculty of Computer Science, Universitas Indonesia, Indonesia, in 2014 and 2015, respectively, and the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2020. He is currently at Facebook AI, USA. Previously, he was a Research Scientist Intern with Google Brain and Facebook AI. His research interests include machine learning, speech recognition, speech synthesis, and natural language processing.



SAKRIANI SAKTI (Member, IEEE) received the B.E. degree *(cum laude)* in informatics from the Bandung Institute of Technology, Indonesia, in 1999, the M.Sc. degree in communication technology from the University of Ulm, Germany, in 2002, and the Ph.D. degree from the Dialog Systems Group, University of Ulm, in 2008. In 2000, she received the DAAD-Siemens Program Asia 21st Century Award to study in the communication technology, University of Ulm. During her thesis work, she worked with the Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. From 2003 to 2009, she worked as a Researcher with ATR SLC Labs, Japan. From 2005 to 2008, she worked with ATR-NICT, Japan. From 2006 to 2011, she worked as an Expert Researcher with NICT SLC Groups, Japan. From 2009 to 2011, she has served as a Visiting Professor for the Computer Science Department, University of Indonesia (UI), Indonesia. From 2011 to 2017, she was an Assistant Professor with the Augmented Human Communication Laboratory, NAIST, Japan. She has also served as a Visiting Scientific Researcher for INRIA Paris-Rocquencourt, France, from 2015 to 2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation." She is currently a Research Associate Professor with NAIST, and a Research Scientist with the RIKEN, Center for Advanced Intelligent Project AIP, Japan. She was actively involved in collaboration activities, such as the Asian Pacific Telecommunity Project, from 2003 to 2007, and A-STAR and U-STAR, from 2006 to 2011. Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition and synthesis, spoken language translation, affective dialog systems, and cognitive communication. She is also a member of JNS, SFN, ASJ, ISCA, and IEICE. She is also the Officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a Board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU).

SATOSHI NAKAMURA (Fellow, IEEE) received the B.S. degree from the Kyoto Institute of Technology, in 1981, and the Ph.D. degree from Kyoto University, in 1992. He was an Associate Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, from 1994 to 2000. He was the Director of the ATR Spoken Language Communication Research Laboratories, from 2000 and 2008, and the Vice President of ATR, from 2007 to 2008. He was the Director General of the Keihanna Research Laboratories and the Executive Director of the Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan, from 2009 to 2010. He is currently a Professor with the Graduate School of Science and Technology, Nara Institute of Science and Technology, a Project Leader of the Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, an Honorary Professor with the Karlsruhe Institute of Technology, Germany, and the ATR Fellow. He is also the Director of the Augmented Human Communication Laboratory and a Full Professor with the Graduate School of Information Science, Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world, including C-STAR, IWSLT, and A-STAR. He received the Yamashita Research Award, the Kiyasu Award from the Information Processing Society of Japan, the Telecom System Award, the AAMT Nagao Award, the Docomo Mobile Science Award, in 2007, and the ASJ Award for Distinguished Achievements in Acoustics. He also received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampolli Award, in 2012. He has been an Elected Board Member of International Speech Communication Association, ISCA, since June 2011, the *IEEE Signal Processing Magazine* Editorial Board Member, since April 2012, and the IEEE SPS Speech and Language Technical Committee Member, since 2013.