Doctoral Dissertation

Multimodal Machine Chain

Johanes Effendi The
Program of Information Science and Engineering
Graduate School of Science and Technology
Nara Institute of Science and Technology

Supervisor: Professor Satoshi Nakamura
Augmented Human Communication Lab.
(Division of Information Science)

Submitted on June 30, 2021

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
DOCTOR of ENGINEERING

Johanes Effendi The

Thesis Committee:

Professor Satoshi Nakamura
(Supervisor, Division of Information Science)
Professor Michiel Bacchiani
(Co-supervisor, Google)
Professor Taro Watanabe
(Co-supervisor, Division of Information Science)
Associate Professor Sakriani Sakti
(Co-supervisor, Division of Information Science)

To the memory of my brother Christian Effendi,
who always believed in my abilities more than I did.

Acknowledgements

"From the fullness of His grace, we have all received one blessing after another."
(John the Apostle)

I want to express my gratitude to Professor Satoshi Nakamura for welcoming me to his lab and supervised me for these five years. He introduced me to the scientific research environment in Japan through an internship and recommended me for the MEXT-IPGP scholarship program. His great insight and leadership have been my inspiration on how I strive to become a good researcher.

I would like to thank Associate Professor Sakriani Sakti for becoming my supervisor and friend through my research here in Japan. I have learned the skills and the best practices of research, writing, and presentation. I am grateful that she motivates me through all the hardships during my time at NAIST.

I also want to thank the thesis committee, Professor Michiel Bacchiani and Professor Taro Watanabe for their valuable comments and suggestions for this thesis. I would like to extend my gratitude to all the faculty I have worked with in the Augmented Human Communication (AHC) Lab., for their support in every aspect of my research life.

I would like to thank the members of AHC Lab., for being such a great companion and delivering a friendly environment during my stay there. My special thanks to our lab secretary Matsuda-san for helping me with various challenges that occur during my stay in Japan; Thanks for taking me to the hospital! I also would like to thank Yoshinaka-san from RIKEN for her assistance from the AIP.

I would like to take the chance to thank Marcello Federico, Roberto Barra-Chicote, and Yogesh Virkar for their guidance and insight during my internship at Amazon Web Service (AWS). It is a very invaluable experience that I received there. I also want to extend my gratitude to Michiel Bacchiani, Yuma Koizumi, and Shigeki Karita from Google Tokyo. I learned a great deal and I am grateful for the warm welcome.

I wish to acknowledge the support of the late Dr. Mirna Adriani, for introducing me to the world of research and encouraged me to continue my study abroad.

Throughout my life, I have received blessings that strengthen and encourage me. I am blessed to have my family and friends that are always giving me support and encouragement, whose warmth can be felt from thousands of kilometres apart.

Multimodal Machine Chain*

Johanes Effendi The

Abstract

Researchers have been working in speech technology for many decades. State-of-the-art automatic speech recognition (ASR) and text-to-speech synthesis (TTS) 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 parallel speech data and its corresponding transcriptions. To improve the performance in the presence of unexpected acoustic variability, we usually collect more data to train more detailed models. Unfortunately, such a method can only be used to train the model for about 10-20 of the world's most common languages. For many others, the parallel data of speech and its transcriptions are usually unavailable, which makes such models hard to implement.

On the other hand, human learning does not rely on parallel data. We can learn from any experience, even if the examples are not provided at the same time. These experiences are perceived in the form of senses, such as auditory and visual, which shares complementary behaviour to ensure flexible learning from any modality (i.e. speech, text, image) in the form of a feedback loop. Inspired by this mechanism, we propose a multimodal machine chain (MMC) as a general framework that accommodates learning in any kind of modality and data availability (i.e. paired, unpaired, single-modality). In this framework, a cross-modal model is able to learn from non-parallel data through feedback it receives after mapping the input into other modalities. Consequently, more modalities, in this case, means more feedback can be made, which therefore enable model learning with fewer data. This makes our proposed learning strategy beneficial for under-resource language, where such technologies matter the most.

*Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, June 30, 2021.

This thesis contribution is four-fold. First, we defined a general framework that enables cross-modal model training in any modality and any data availability. Second, we showed that our MMC framework can be used to enable semi-supervised cross-modal collaboration that allows learning from a single-modality data, which modality is unrelated. Third, we pushed the level of supervision boundary into weakly-supervised learning, to enable a speech-to-text mapping using a visually-connected non-parallel data. Finally, we showcase our proposed MMC framework capability to learn a self-supervised discrete speech representation to enable image-to-speech generation without text. All these four contributions in the form of MMC framework and its applications shows its capability to enable speech processing model learning for low-resource language or even unknown untranscribed language.

Keywords:

semi-supervised learning, weakly-supervised learning, self-supervised learning, speech recognition, multimodal information processing, low-resource language

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is commonly known as the most natural means of communication. Speech communication consists of listening and speaking activities to perceive and convey information. We perceive information by processing the speech signal with the help of the ear as the organ that processes auditory senses and passes it to the brain. On the other hand, we convey information by uttering the speech that we are thinking with our vocal tract. Both of these activities are coordinated by the brain that maintains the structure of the perceived and conveyed information in the form of language. Although we mostly communicate by speech, visual modality also has an important role to support communication. We see an object with our eyes, and then our brain recognizes it. Then, we can describe it in the language we know in the form of speech or textual information. In conclusion, such triangle modalities of **visual, auditory, and textual** are the principal medium of human communication.

1.1. Human Speech Communication

1.1.1 Human Speech Chain Perspective: Speech Production and Perception

Denes et al. described the relationship between human listening and speaking activities in the form of human speech chain [1]. Although listening and speaking activities are done by a different organ, both these activities are closely related.

Spoken messages are propagated from the speaker's mind to the listener's mind (Figure 1.1). 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. The relationship between these two activities is so close that children who lose their hearing often have difficulty in producing a clear speech because they are unable to monitor their speech [8].

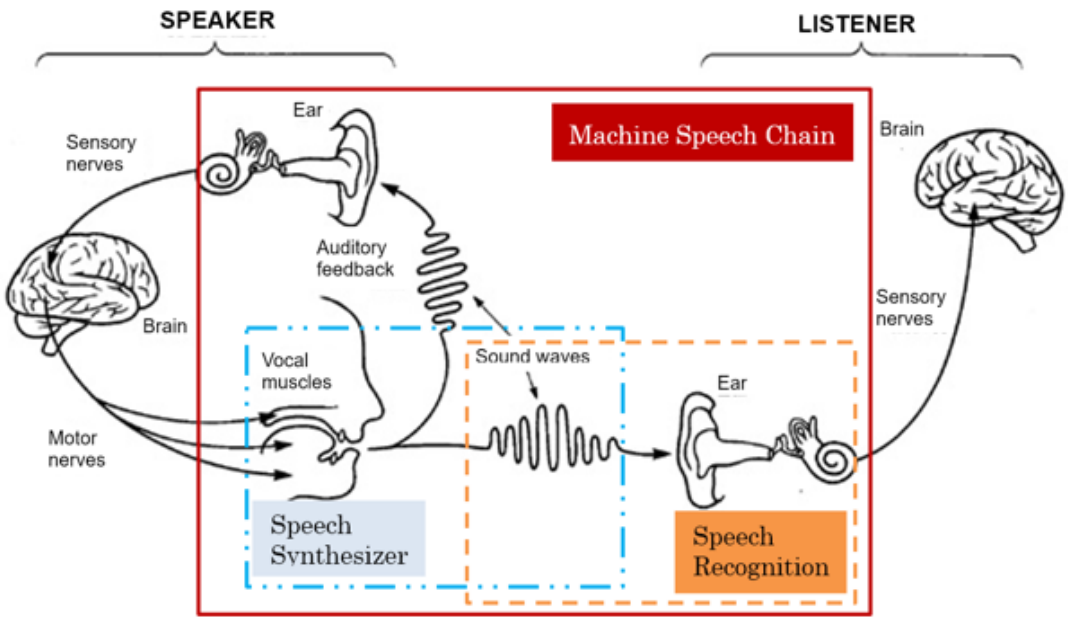

Figure 1.1: Human speech chain [1]

1.1.2 Multimodality and Flexibility in Human Speech Communication

Visual modality as part of human communication

Although most communication is mainly conveyed through speech and text modality, visual modality is also often used alongside them. Figure 1.2 illustrates what is happening when someone hears a speech of "a dog is running" while looking at a running dog which represents the speech message. The visual modality seen

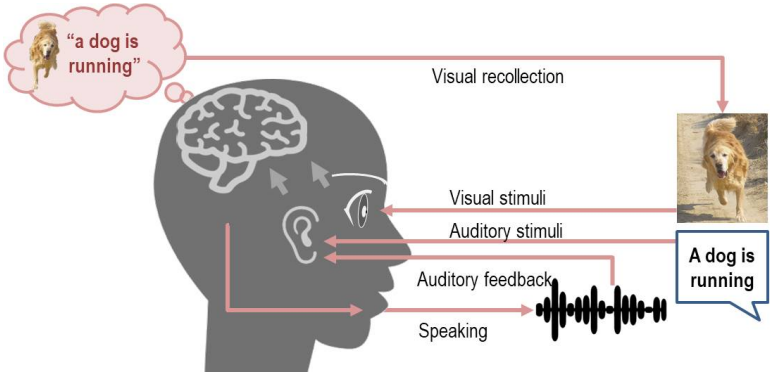

Figure 1.2: Multimodal perception in human speech communication

by the eyes supply information about colour, texture, and other visual aspects of the viewed object, which helps us perceive what we see [9] by enriching the speech we listened to. There is evidence that the heard speech and the viewed scene are perceived altogether in the form of cross-modal processing as an audio-visual speech [10]. In addition, visual modality can also complement the missing information when it is difficult to infer from auditory channels [11]. Given these points, we can conclude that visual information is also a crucial part of speech communication.

Learning flexibility

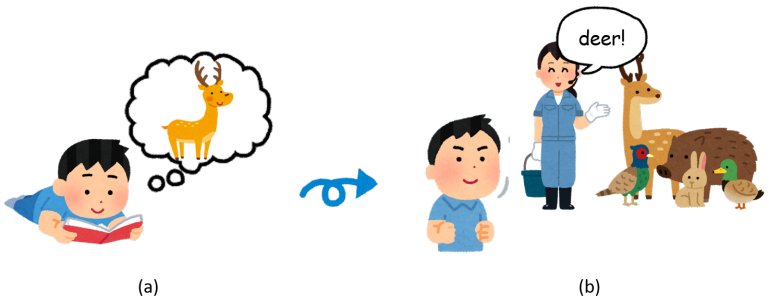

Figure 1.3: Illustration on human learning flexibility. (a) learning from visual+text, and (b) learning from speech+text.

Figure 1.3 illustrates human learning flexibility. Assumes that on the left

side (a), a child reads a picture book with an image of a deer. Then, later in a zoo, a guide explains verbally that the present animal is a deer. That child can easily recall what he has learned from the picture book. Now the information of how to pronounce "deer" (auditory) and how to write "deer" (textual) are both linked by the visual modality (i.e. how a deer looks like). From this example, we can conclude that in order to learn something from experience, a human does not need both information provided at the same time. Even when one modality is not present, we can still learn and improve any of our communication skills. This flexibility is possible because of the multimodality in human communication, where each modality shares complementary behaviour.

1.2. Technology for Speech Communication

For years, machine learning has been trying to mimic human speech communication by automating the task that human does. One of the most commonly investigated is the cross-modal task, which is an attempt to automate the mapping of one modality to another. Consequently, a machine learning model that attempts to map one modality to another is called a cross-modal model. In speech processing, there are two main cross-modal tasks to mimic human communication ability, which are speech recognition (to simulate listening) and speech synthesis (to simulate speaking).

1.2.1 Speech Recognition

Figure 1.4: ASR Model

Automatic speech recognition (ASR) is a speech technology that attempts to transcribe a speech into a text transcription (Figure 1.4). This process automates human perception of listening, where human recognizes speech that she is listening to, and text transcription represents the content of the speech in the form of phoneme or grapheme (i.e. character, word).

There are several major approaches since the earlier times, which follows the increase of computation power [12]. At the late 1960s, feature-extraction algorithms that enables fast Fourier transform (FFT) [13] and dynamic time warping (DTW) [14] were developed. Then, these features are proven to be useful since the development of hidden Markov models (HMM) for speech processing [15, 16]. After that, the use of Gaussian mixture models (GMMs) to model the acoustic information of speech, while HMM is used for the phonetic sequence were widely adopted by the speech community in the 1990s. This has been proven to widen the application of the ASR task to recognize a larger vocabulary size, with speaker-independent condition [17].

From the neural network side, there was also the first usage of convolutional networks for speech in the form of a time-delay neural network (TDNN) [18]. The use of a hybrid HMM/MLP architecture has also been used [19, 20], although these attempts were limited by computational resources until several decades later with the rise of GPU for a deep neural network. This contributed to the rise of end-to-end deep neural network for ASR which replaces GMM for acoustic modelling with recurrent neural network (RNN) [21], convolutional layers [22], or self-attention encoder [23]. Recently, deep learning-based state-of-the-art ASR frameworks have even been shown to reach human parity in performance [24, 25].

1.2.2 Speech Synthesis

Figure 1.5: TTS Model

A text-to-speech (TTS) is a speech technology that aims to synthesize speech given text input (Figure 1.5). Jurafsky and Martin (2009) describes three early paradigms for TTS: articulatory synthesis, concatenative synthesis, and formant synthesis. Articulatory synthesis models the physics of the vocal tract as an open tube, in order to synthesize speech [26, 27, 28]. Concatenative synthesis combines several speech units in the form of diphones, to synthesize a speech utterance considering the F0, stress, duration, and formant distance between neighbour-

ing units [29, 30]. The last paradigm is formant synthesis, which attempts to synthesize a spectrogram similar to the reference speech [31, 32].

The recent rising interest in the deep learning approach fosters the development of several sequence-to-sequence models which falls under the formant synthesizer paradigm. A model such as Tacotron [2] aims to synthesize a mel-spectrogram given phoneme sequence as an input, with an attention mechanism to bridge the encoder and decoder. FastSpeech [33, 34] and Parallel Tacotron [35, 36] proposed a non-autoregressive approach, by replacing attention mechanism with a duration prediction mechanism.

1.2.3 Current Limitations and Our Proposal

Problems:
Widely-used supervised learning requires large amount of parallel data.

The last two sections have described the speech researchers attempt to simulate human speech communication skills of listening and speaking into a machine. Based on the metrics that are used by them to evaluate their models, the current state-of-the-art ASR systems have been known to successfully reach parity with humans [24, 25]. Although it is reassuring, such models can only be used to perfectly recognizing the speech of the top 10-20 of the world's most common languages. It is difficult to be applied for many other languages because the required speech and the corresponding transcriptions are usually unavailable.

The problem lies in the learning paradigm used to train such state-of-the-art models. 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, they usually collect more data to train a more detailed model. Therefore, a novel learning mechanism to reduce the need for parallel data to train a speech processing model is needed, especially for an under-resourced language.

Moreover, there has been some attempt to incorporate more modalities in a system, inspired by human communication multimodality. For example, there has been some work in audio-visual ASR [37, 38], multimodal machine translation [39, 40], and visual TTS [41, 42] to incorporate visual modalities in a natural lan-

guage model. Several advantages have been reported such as reduced ambiguity, improved performance, and a more natural result.

However, such research direction will introduce a new kind of limitation, which is related to the curse of dimensionality [43]. Similar to the previous problem, such models are trained in a supervised manner, which heavily relies on the parallel dataset. Therefore, the more modality we add to the system, the more difficult it is to get the parallel data of all those included modalities.

Existing approaches:
Solutions for specific modalities and specific data condition.

The most common way to solve the limited data problem is by using semi-supervised learning. Label propagation is the first attempt at learning from partially unlabeled data [44, 45]. First, a model is trained with the labelled portion of the data. Then, the trained model generates a pseudo label for the unlabeled portion of the data, so that it can be used to continue the training.

Similarly, a dual learning mechanism is inspired by this method that enables learning from source-to-target by feedback links, which provide the possibility of training models with unpaired datasets. He et al. [46] proposed dual learning in neural machine translation (NMT). In their work, a source-to-target NMT model (primal) receives feedback from the target-to-source NMT model (dual). This collaboration between two agents of primal and dual enables learning from monolingual and unpaired data. In addition, there are also some similar approaches in image processing such as DiscoGAN [47], CycleGAN [48], and DualGAN [49]. Although this method looks promising, both the primal and dual agents are within the same modality, which diverges the task from the reality of human multimodal communication.

For a cross-modal task, inspired by the closed-loop speech chain in human communication (Section 1.1.1), a machine speech chain framework is proposed to enable ASR and TTS model training from unpaired dataset [50, 51, 52, 53]. Machine speech chain integrates human speech perception and production behaviours 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 (Figure 1.6).

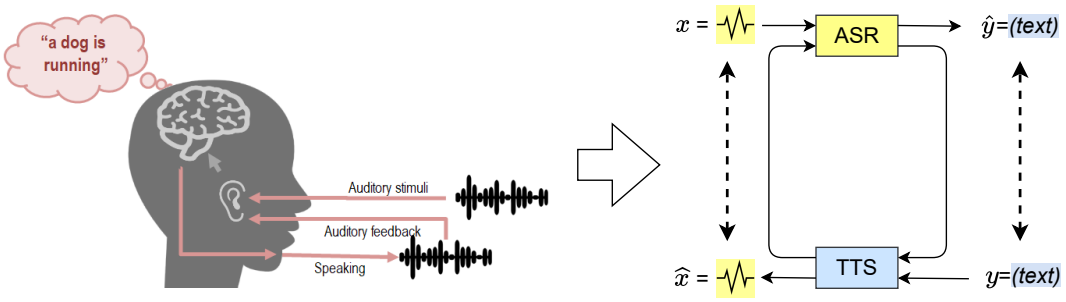

Figure 1.6: Machine speech chain (right) inspired by human speech chain (left) as an attempt to enable semi-supervised learning from unpaired data.

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. The speech-only data is transcribed by a pretrained ASR model, which text hypothesis then can be used by pretrained TTS model to generate a speech hypothesis. Therefore, the TTS model can be updated by the speech reconstruction model. In reverse, a speech hypothesis can be synthesized from text-only input, which then transcribed by the ASR model. Using the reconstruction loss of the original text and the text hypothesis, we can update the ASR model.

Nevertheless, in the cases described before, all of the learning strategies are designed specifically to handle a modality in their particular task. For example, although the machine speech chain reduces the need for parallel data, the unpaired dataset being used is still related to the input and output of the cross-modal model itself. Furthermore, in the "Watch, Listen, Attend, and Spell" audio-visual ASR framework, they arranged flexible learning due to an imbalanced number of speech and lip video data. However, such a method can only be used in their specific multi-source architecture.

Although these approaches are indeed reducing the need for parallel data to some extent, the idea cannot be used in general for other modalities and other model architecture due to different characteristics. We propose that there should be a general framework that defines learning strategies that can be applied for any modalities and any kind of model architecture. Consequently, that general

framework should be capable to enable learning for such data conditions with any level of supervision.

Our Proposal:
General framework to enable learning with any data condition by cross-modal collaboration.

We propose that the solution to these problems are actually can be learned from how humans learn to communicate in their early days which does not always need parallel data. Young children can communicate effectively with their parents in their native language even before they learn to read or write [54]. This means that if later those children can transcribe a speech, that does not mean they need a special learning time where both the speech utterance and the text transcription are given at the same time, as described in Section 1.1.2. From the example in Figure 1.3 we can also see that the learning process is not always happening when all of the materials are presented together, but also through a confirmation or feedback in an interaction.

Therefore, a system that attempts to simulate human communication must be able to learn from any modalities, even when they are not parallel (at the same time). In addition, the system should also be able to learn not only from a supervised approach but also from some confirmation or feedback. Looking at the dual learning or the machine speech chain mechanism, the feedback comes from the closed-loop mechanism. While doing a set of cross-modal operations in a loop, the speech chain can generate a reconstruction loss, which can be counted as a feedback mechanism that helps the cross-modal model in the loop to learn.

This thesis focuses on generalizing this idea into any modality, so that the more modality we have, the more feedback we can generate through cross-modal mapping in between them. Therefore, each of the cross-modal models in the framework can collaborate in the form of a closed-loop chain. Consequently, we also need to ensure that the feedback can be yielded and be used for any kind of cross-modal model architecture. This idea addresses the low-resource data limitation by also replacing the multimodality limitation as an opportunity. In this thesis, we propose a multimodal machine chain that is inspired by this idea using three kinds of modalities: speech, text, and images; while also generating

feedback in the form of reconstruction loss, given various data availability (i.e. paired, unpaired, single modality).

1.3. Thesis Scope

1.3.1 Thesis Contribution

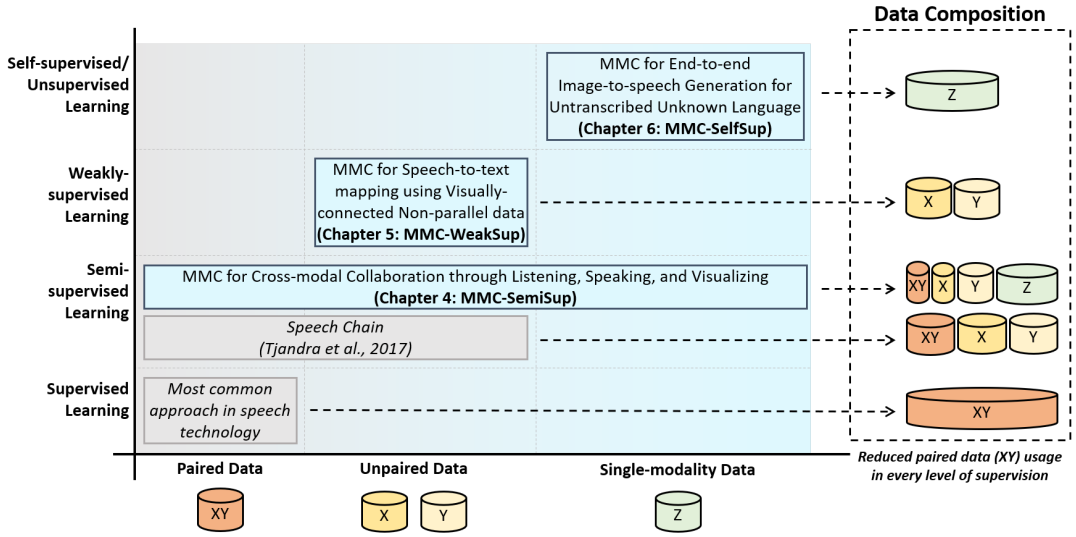

Figure 1.7: This thesis contribution overview, x-axis: data conditions, y-axis: supervision levels.

In this thesis, we propose a generalization of the chain mechanism, to enable cross-modal model learning from any kind of data availability. Figure 1.7 shows the realm of the problem in enabling speech processing model training in any data condition (x-axis), with the supervision levels as the implication (y-axis).

First, we formally define the general framework for multimodal machine chain (MMC). We designed the MMC framework to enable model training from any modality and any data availability. As a consequence, the MMC framework consists of learning strategies in various levels of supervision.

Second, we show the MMC framework generalization effectiveness with an experiment to improve a cross-modal model by leveraging data from unrelated modality. This approach enables semi-supervised learning from the combination

of paired data, unpaired data, and unrelated single modality data. We named this approach as **MMC-SemiSup**.

Third, to show the robustness of the MMC framework in terms of data availability, we attempt to realize speech-to-text mapping using a visually-connected non-parallel data. Therefore, we showcase that the MMC framework can also enable learning even when the data is fully unpaired. We label this approach as **MMC-WeakSup**.

Finally, we tackle one more cross-modal learning problem in terms of representation. Our proposed MMC framework not only enables training from different modalities but also able to create a better representation when the optimal one is not available. We call this as **MMC-SelfSup** for this self-supervised approach to create a discrete speech representation. We investigate it in the Image2Speech task, where we attempt to learn a speech representation that enables end-to-end image-to-speech generation without using any text.

In this thesis, we limit our modality scope to textual, visual, and auditory. We use speech data to represent auditory modality, image data for visual modality, and the transcription or caption of these data as a textual modality.

1.3.2 Thesis Outline

The structure of the remaining chapters of this thesis is as follows. Chapter 2 is an introduction to the language technologies in various modalities. We describe what kind of neural network-based cross-modal model that we use to showcase our proposed MMC framework in the subsequent chapters. Chapter 3 describes several levels of supervision in machine learning and the formal definition of the MMC framework to tackle each of those levels of supervision.

Then, Chapter 4 describes our attempt to use our MMC framework for cross-modal collaboration through listening, speaking, and visualizing. We describe our attempt to enable semi-supervised learning of speech processing model with either paired, unpaired, and unrelated modality data.

Chapter 5 covers our attempt to use the MMC framework to enable speech-to-text mapping, even when the data is just weakly-connected. This shows the robustness of our framework to realize weakly-supervised learning, even when the data condition is more extreme than before.

Chapter 6 tackles the representation problem in realizing a text-free image-to-speech generation, which is beneficial for untranscribed unknown language. We use our MMC framework here to learn an optimal representation for speech, to enable end-to-end learning directly from image modality to speech modality.

Finally, we conclude our thesis in Chapter 7. In addition, we also discuss further possible future research in the topic related to this thesis.

Chapter 2

Language Technologies in Various Modalities

This chapter describes the current state-of-the-art cross-modal model commonly uses in the modalities covered in this thesis.

2.1. Neural Speech-Text Processing Models

2.1.1 Automatic Speech Recognition (ASR)

Commonly known sequence-to-sequence ASR model resembling the Listen, Attend, and Spell (LAS) framework, uses location-aware attention [21]. As illustrated in Figure 2.1, this model encodes a speech feature (i.e. Mel-frequency cepstrum (MFCC), mel-spectrogram) $\mathbf{x} = [x_0, \ldots, x_n]$ 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. The encoder architecture is usually pyramidal in its depth to reduce the length of the input.

Then, the decoder receives input character y_i for timestep i, which is converted into decoder hidden state d_i^{ASR}. Then, to condition the generation process against the encoder state, an attention mechanism is generating a context vector c_t by creating a weighted sum of the encoder states e_s^{ASR}, given the current decoder

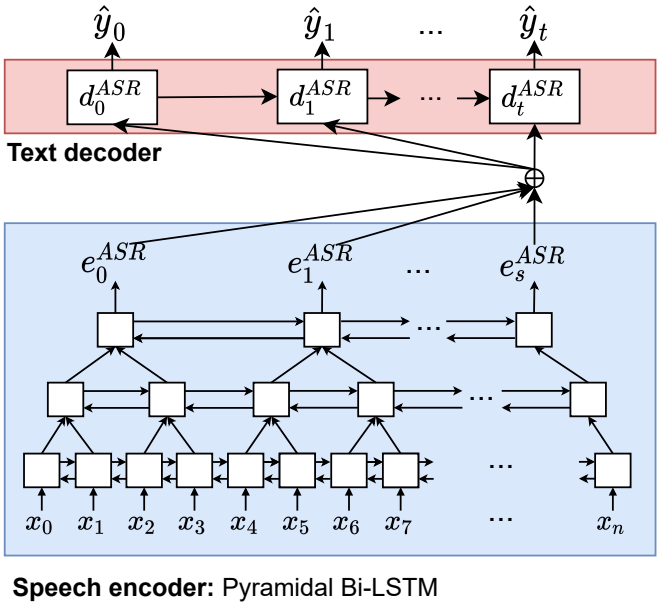

Figure 2.1: ASR model with pyramidal Bi-LSTM

hidden state d_i^{ASR} as follows:

$$c_t = \sum_{s=1}^{S} a_t(s) * e_s^{ASR} \tag{2.1}$$

$$\begin{aligned} a_t &= \text{Align}(e_s^{ASR}, d_i^{ASR}) \\ &= \frac{\exp(\text{Score}(e_s^{ASR}, d_i^{ASR}))}{\sum_{s=1}^{S} \exp(\text{Score}(e_s^{ASR}, d_i^{ASR}))}, \end{aligned} \tag{2.2}$$

while variations [55] for score function includes:

$$\text{Score}(e_s^{ASR}, d_i^{ASR}) = \begin{cases} \langle e_s^{ASR}, d_i^{ASR} \rangle, & \text{dot product} \\ e_s^{ASR\intercal} W_s d_i^{ASR}, & \text{bilinear} \\ V_s^{\intercal} \tanh(W_s[e_s^{ASR}, d_i^{ASR}]), & \text{MLP} \end{cases} \tag{2.3}$$

Then, the hypothesis probability can be produced by an output layer $p_t = out([c_t, d_t^{ASR}])$. The loss can be calculated as a softmax cross-entropy loss between the hypothesis probability p_{i+1} and the one-hot vector of the next character y_{i+1}.

The loss function for ASR model can be formulated as:

$$L_{ASR} = -\frac{1}{t}\sum_{i=1}^{t}\sum_{c=1}^{C} \mathbb{1}(y_{i+1} = c) * \log p_{y_{i+1}}[c]. \tag{2.4}$$

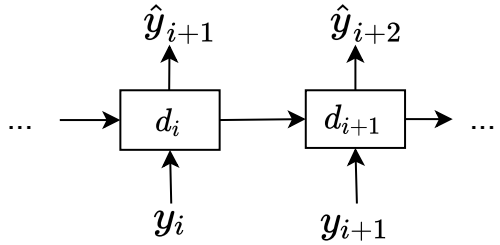

Figure 2.2: Teacher forcing in autoregressive model. Regardless of the predicted token $\hat{y}_{i+1}$, the original y_{i+1} is being used as the input for the next timestep.

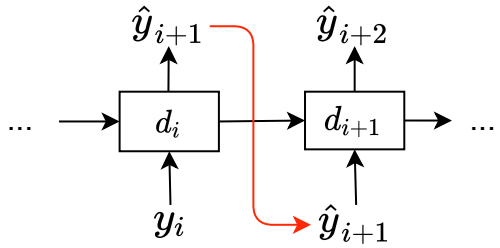

Figure 2.3: Prediction step without teacher forcing in autoregressive model. The predicted token $\hat{y}_{i+1}$ is being used as the input for the next timestep (red line).

Teacher-forcing is used during training, which means that for the next timestep $i + 1$, y_{i+1} is being used as the input for the decoder (Figure 2.2). During inference, the one-hot label of a class with the highest probability in timestep i is being used as the input for the next timestep (Figure 2.3).

2.1.2 Text-to-speech Synthesis (TTS)

A sequence-to-sequence TTS receives a text utterance $\mathbf{y} = [y_0, .., y_s]$ and learn to generate a speech feature $\mathbf{x} = [x_0, .., x_t]$ by optimizing its parameters. The most common model is the Tacotron TTS [2]. Input sequence consisting of characters

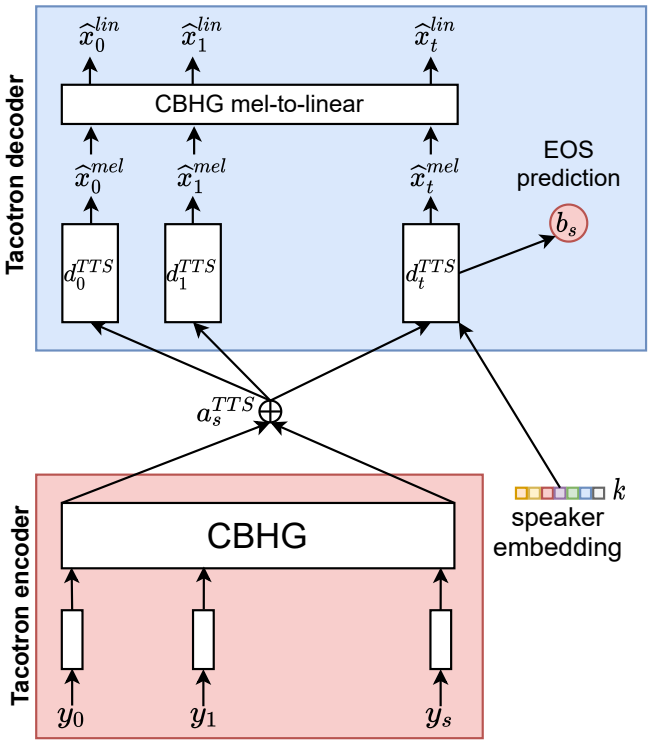

Figure 2.4: TTS model resembling the Tacotron[2] architecture. In a multispeaker setting, the decoder is conditioned by k, which is the speaker embedding input.

or phonemes are commonly used for TTS input. First, the inputs are projected into vectors by embedding layer. Then, it is projected to the CBHG block (1D **C**onvolution **B**ank + **H**ighway + bidirectional **G**RU) with eight filter banks (filter size from 1 to 8) which produces encoder state $e^{TTS} = [e_0^{TTS}, e_1^{TTS}, \ldots, e_s^{TTS}]$.

CBHG (See Figure 2.5) is a module that is commonly used in a text-to-speech model such as Tacotron. It starts with a 1D convolution bank with the stacking of several filters. The width of the filters is within the range of 1 to K, so that there are various ranges when encoding the inputs. After stacking the filter result together and max pooling, it is inputted into a 1D convolution to preserve the time dimension. Then, the residual connection is added with the latter representation to make a residual connection. Then, the hidden representation is inputted into a highway layer so that the output can be encoded further by the bidirectional LSTM or GRU to get a high-level representation of the sequence.

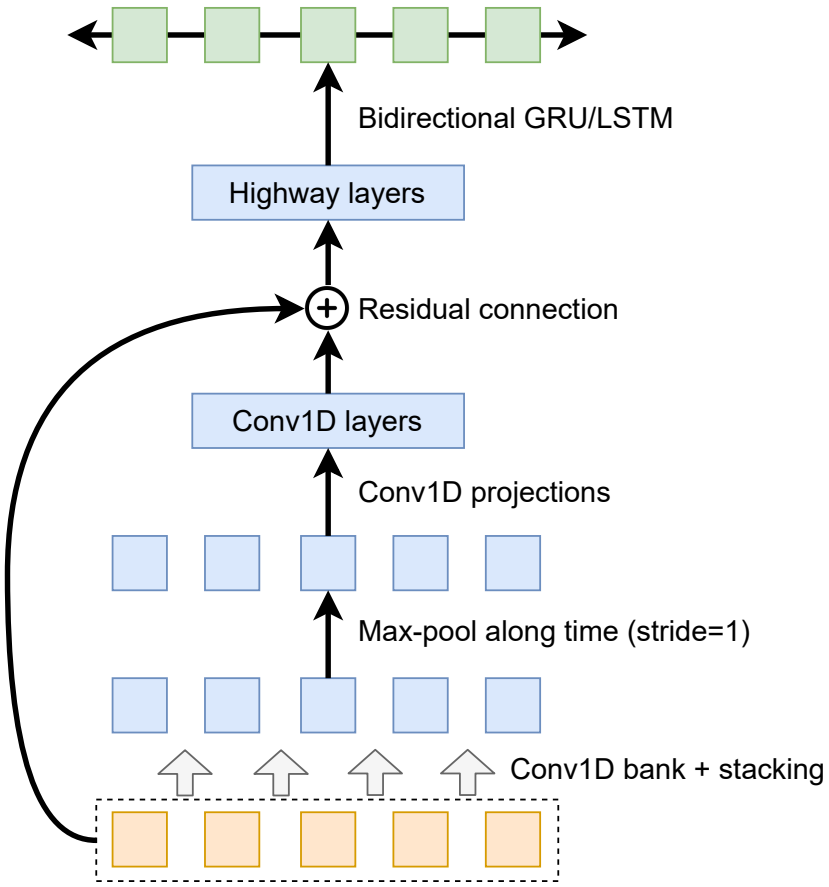

Figure 2.5: Illustration of a CBHG layer. A stacking of several 1D convolution enable narrow and wide context range.

In addition, for a multispeaker model, a speaker embedding is also used to condition the generated speech. The speaker embedding k is inputted to the speaker embedding layer as-is or a speaker id can be inputted. In a condition where no speaker information is available to generate the speech, the speaker embedding k can be randomly sampled from a known distribution.

The decoder part is separated into two steps. The first step is to generate mel-spectrogram frames $[\hat{x}_0^{mel}, \hat{x}_1^{mel}, \ldots, \hat{x}_t^{mel}]$ and an end-of-speech prediction $b_s \in [0, 1]$. If the current frame i is the end of speech, then the value of b_s is 0, else 1. The generation is conditioned to the encoder states using attention mechanism, similar to ASR (Section 2.1.1). Finally, a CBHG mel-to-linear is used to project

the generated mel-spectrogram into a 1025 dimensional linear spectrogram.

For training the first step, an $L2^2$ loss is used to compare the generated mel-spectrogram with the target mel-spectrogram such as:

$$L_{mel} = \frac{1}{t}\sum_{i=0}^{t} \|x_i^{mel} - \hat{x}_i^{mel}\|_2^2. \tag{2.5}$$

This loss function is also used in the next step for projecting mel-spectrogram to the linear spectrogram, so that:

$$L_{lin} = \frac{1}{t}\sum_{i=0}^{t} \|x_i^{lin} - \hat{x}_i^{lin}\|_2^2. \tag{2.6}$$

In addition to that, a binary prediction loss is used for the end-of-speech prediction as

$$L_{EOS} = b_s \log(\hat{b}_s) + (1 - b_s)\log(1 - \hat{b}_s). \tag{2.7}$$

Finally, all the losses are summed to train the TTS model, so that:

$$L_{TTS} = L_{mel} + L_{lin} + L_{EOS} \tag{2.8}$$

During the inference, the phase spectrogram can iteratively be estimated using Griffin-Lim [56] algorithm from the linear spectrogram, and then reconstructed with the inverse short-time Fourier transform (STFT) to produce the speech signal. Inversion to waveform can also be done with other parametric vocoders such as WaveNet [57] or Universal Vocoder [58].

2.1.3 Quantization of Speech Features (VQ)

An autoencoder [59] ensures that the system reconstructs output that is most similar to the input data it is given. The architecture of an autoencoder usually consists of an encoder that encodes the input into a compressed representation and a decoder that reconstructs the representation. To ensure the statistical properties of the representation's latent space, a variational autoencoder (VAE) [60] was proposed along with training regularization.

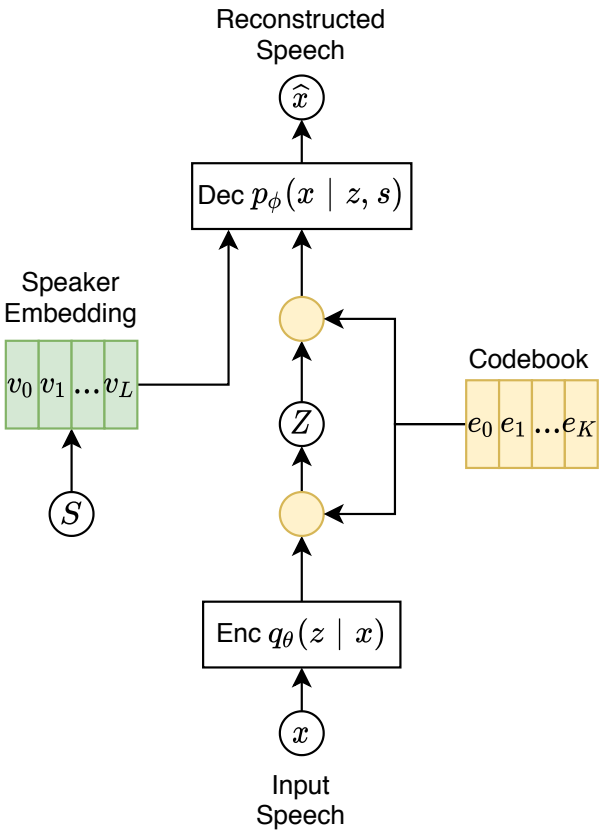

Figure 2.6: VQ-VAE model with speaker embedding

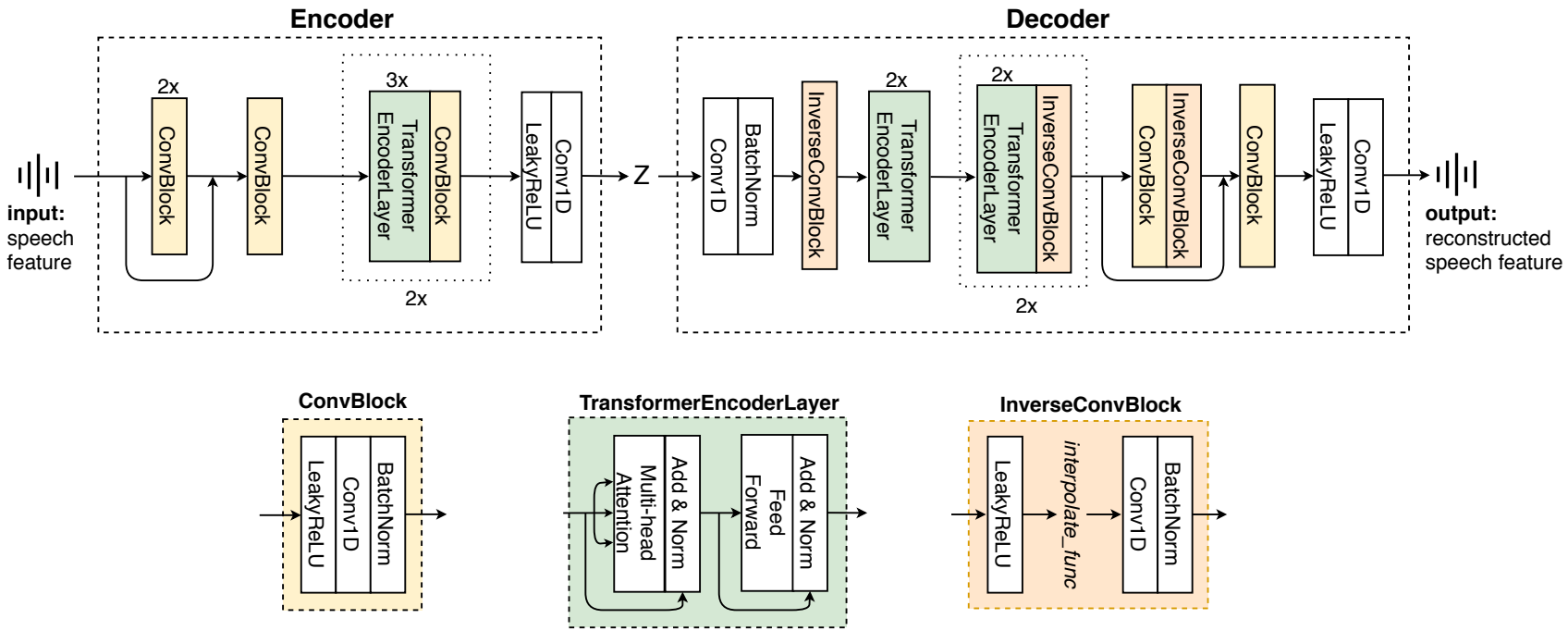

Figure 2.7: Detailed structure of our VQ-VAE model, all layer parameters are similar to Tjandra et al. (2020) [3]

A vector-quantized variational autoencoder (VQ-VAE) [57] is a variant of variational autoencoder (VAE) architecture that generates a discrete latent representation instead of a continuous representation in VAE. In the quantization process, VQ-VAE conditions the latent representation to become the element of the closest code. In this way, we can view the VQ-VAE codebook as a collection of clusters (codes), where continuous representation derives from the mean vector of each cluster (code vectors). The purely discrete representation can also be obtained by simply using the label of the cluster that an encoded representation belongs to.

As shown in Figure 2.6, a VQ-VAE model encodes input x, which is a speech feature such as Mel-frequency cepstral coefficients (MFCC) or Mel-spectrogram. The stack of encoder layers produces an intermediate continuous representation $z \in \mathbb{R}^{D_c}$. Then, it is compared with all possible code vectors in the codebook to find the code with the closest distance between z and one of the possible code vectors in codebook $[c_1, c_2, \ldots, c_K]$. A speaker id s, represented by speaker embedding $V = [v_1, v_2, \ldots, v_L] \in \mathbb{R}^{L \times D}$, is used as an additional condition for the decoder so that the speech reconstruction process $(x|z, s)$ is conditioned on the codebook vector representation c and speaker information s.

Here, the training objective is defined as follows:

$$L_{VQ} = -\log p_\phi(x|z, s) + \|\text{sg}(z) - C\|_2^2 + \gamma\|z - \text{sg}(C)\|_2^2, \tag{2.9}$$

where function $\text{sg}(\cdot)$ stops the gradient, defined as:

$$x = sg(x); \quad \frac{\partial\, \text{sg}(x)}{\partial\, x} = 0. \tag{2.10}$$

There are three terms for a loss L_{VQ}. The first term is a reconstruction loss as a negative log-likelihood. This loss ensures that both encoder and decoder can produce a good speech feature reconstruction $\hat{x}$, as close as the speech feature input x, given latent representation z and speaker information k. The second term $\|\text{sg}(\hat{z}) - C\|_2^2$ is used to update the codebook C so that it is closer to the encoded representation z. In this term, the codebook is updated while the encoder remains the same. Finally, the third term $\|z - \text{sg}(C)\|_2^2$ updates the encoder so that it produces a representation close to the codebook C. In this term, the

codebook remains the same while the encoder is updated. The resulting loss by this term is scaled by a γ coefficient.

The codebook creation process in this study's VQ-VAE model is frame-based. Through several strided convolutions in the encoder part, one code can represent several speech feature frames. For example, if the encoder has three convolution blocks with $[2, 2, 3]$ stride each, then each code in the codebook represents $2 \times 2 \times 3 = 12$ frames. A wider stride means more information is compressed into each code, which makes the representation more robust but harder to reconstruct.

Figure 2.7 shows the detailed schema of our VQ-VAE structure. The encoder-decoder part of our VQ-VAE is similar to that of Tjandra et al.'s Transformer VQ-VAE [3] for unsupervised unit discovery. First, the speech feature input is passed through a residual connection of several convolutional blocks (ConvBlock). Then, the output is passed through multiple layers of multi-head attention (TransformerEncoderLayer). The discrete representation (Z) can be inverted back to a speech feature through a series of interpolation and convolution operations (InverseConvBlock) and multi-head attention layers.

2.2. Neural Image-Text Processing Models

2.2.1 Image Captioning (IC)

An attention-based image captioning model encodes image **z** into high-level features $[e_0^{IC}, \ldots, e_s^{IC}]$ (Figure 2.9). These features are then used as the context for an attentional text decoder to generate hypothesis captions. To get these two-dimensional high-level features, a partial image classification model is usually used by taking the two-dimensional hidden representation after a series of convolution layer (See Figure 2.8).

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} (Similar to equation 2.1). 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

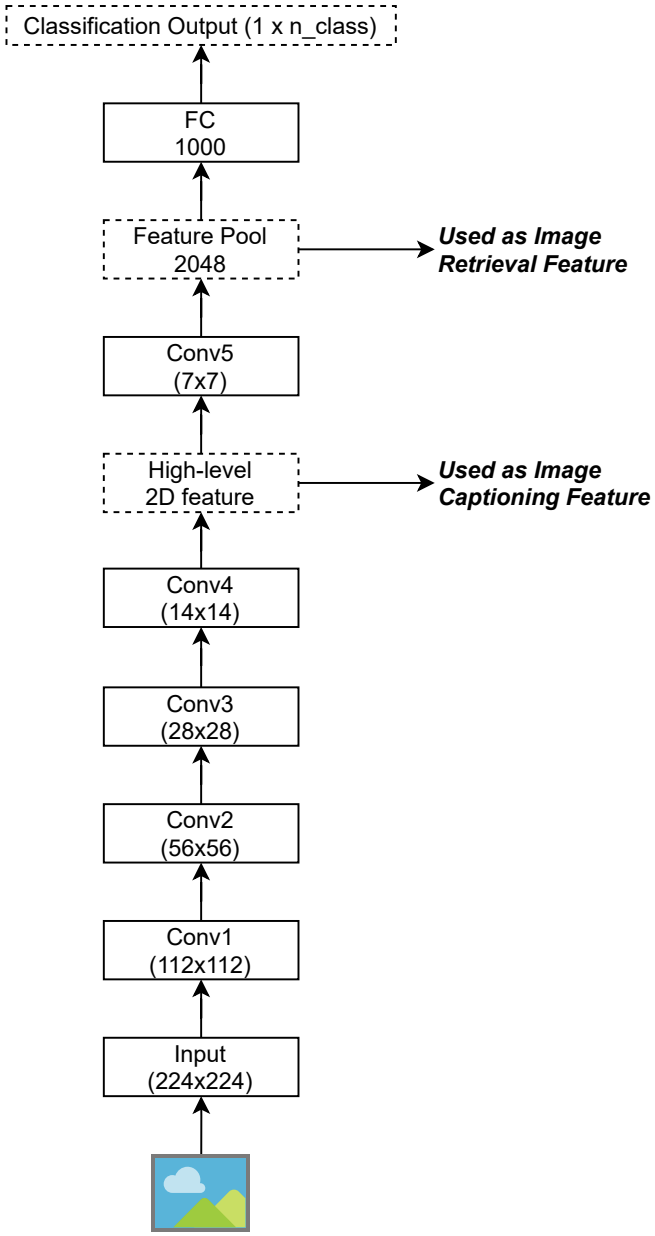

Figure 2.8: ResNet-50 architecture [4] for ImageNet classification task. Dotted box means hidden representation and straight-line box represents neural network layer.

against the original text sequence. ResNet [4] is commonly used as image encoder, and LSTM decoder is commonly used as text decoder, resembling similar

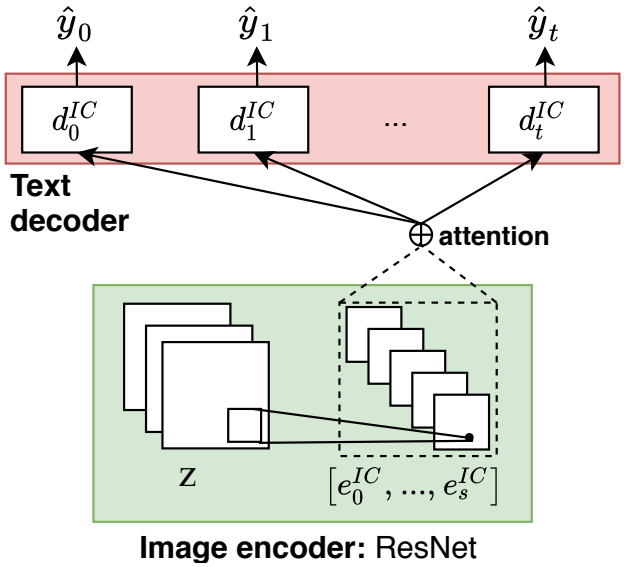

Figure 2.9: IC model with attention mechanism.

architecture with Xu et al. (2015) who proposed the "Show, Attend, and Tell" model [61]. The loss function for IC model for every token at time $t = i$ can be formulated as:

$$L_{IC} = -\frac{1}{t}\sum_{i=1}^{t}\sum_{c=1}^{C} \mathbb{1}(y_{i+1} = c) * \log p_{y_{i+1}}[c], \tag{2.11}$$

which is the cross-entropy loss between the label of the next timestep y_{i+1} and the probability of the predicted token p_{y+1}.

On the other hand, similar to the LSTM-based image captioning model, ResNet [4] can also be used as an image encoder in this type of image captioning model. However, the differences lie in the decoder part. Transformer-based text decoder to generate text captions can be used with the encoded image representation as to the generation condition.

The text decoder is trained using teacher forcing on the sequence of text caption. Commonly known architecture using multi-head attention for the decoder part is the Vaswani et al.'s Transformer model [62], as illustrated in Figure 2.10. It is composed of multiple layers, with three sub-layers for each layer. The first sub-layer is the masked multi-head attention of the target-to-target attention.

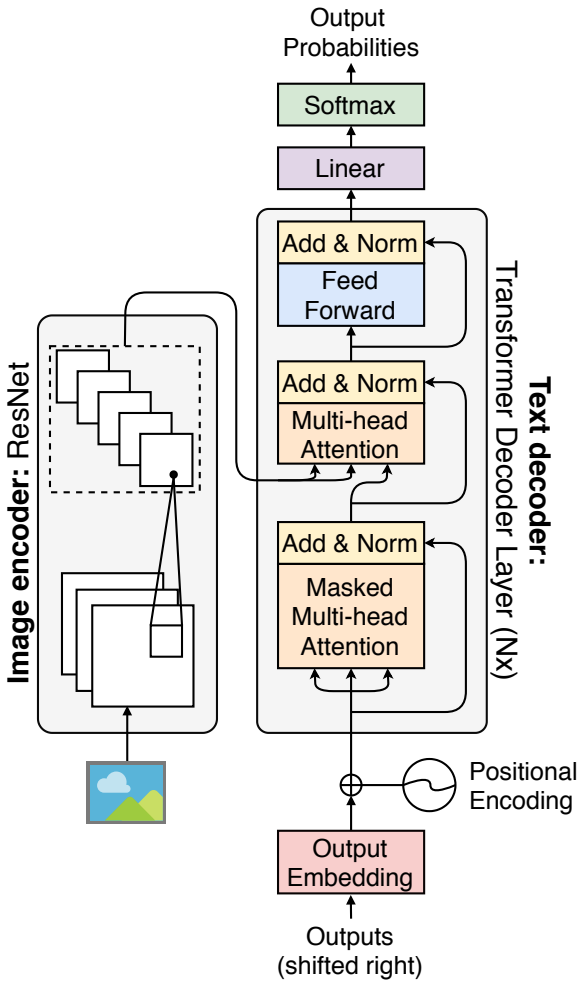

Figure 2.10: Transformer-based image captioning model.

Then, the second layer is the source-to-target multi-head attention. These multi-head attentions resemble a vector query and a set of key-value vector pairs to the output. Finally, the third sub-layer is the position-wise fully connected feed-forward network. After the repetition of these layers, the network is closed with a linear layer having the same size as the vocabulary so that the probability of the next token can be decided.

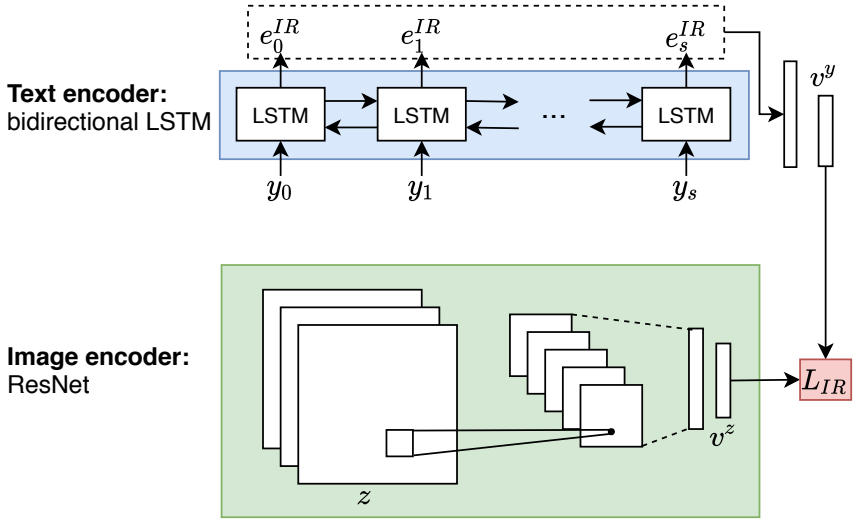

Figure 2.11: IR model.

2.2.2 Image Retrieval (IR)

An image retrieval model encodes image **z** and text caption **y** into embedding vectors v^z and v^y (Figure 2.11). The image encoder is usually constructed by a series of pre-trained convolutional neural networks, followed by pooling and linear transformation at the end to produce image embedding v^z. The vector representation can be taken after the pooling operation in ResNet-50 [4] (See Figure 2.8). The 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, a ranking loss is used with distance d that defines the distancing between positive (v^y, v^z) and negative samples $(\hat{v^y}, \hat{v^z})$. Pairwise rank loss as one of the common loss for image retrieval L_{IR} is defined as follows:

$$\begin{aligned} L_{IR} = & \sum_{|v^y|} \sum_{|\hat{v}^z|} max\{0, M + d(v^y, v^z) - d(v^y, \hat{v}^z)\} + \\ & \sum_{|v^z|} \sum_{|\hat{v}^y|} max\{0, M + d(v^z, v^y) - d(v^z, \hat{v}^y)\} \end{aligned} \tag{2.12}$$

2.2.3 Image Generation (IG)

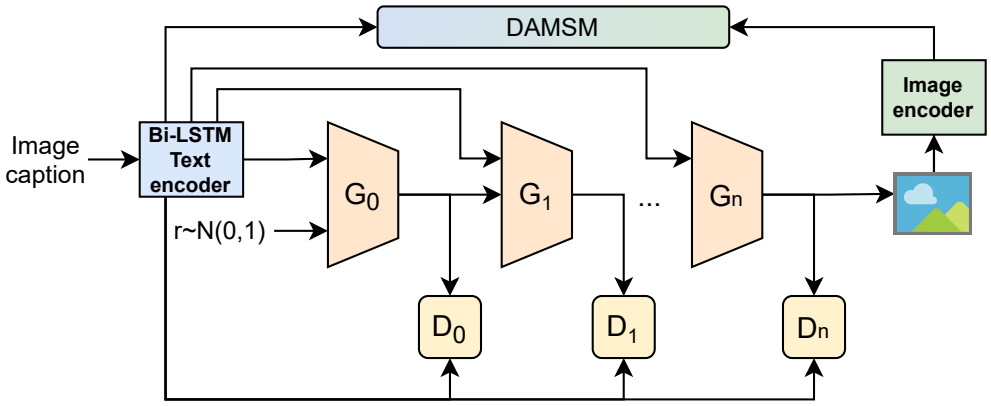

Figure 2.12: AttnGAN model with DAMSM loss.

Generative adversarial network (GAN) model architecture is commonly used to generate an image, given a text caption. GAN is a combination of two networks:

- A **generator** G produces data given a noise sampled from a standard normal distribution. In this image generation task, the generator receives the text caption as an additional condition to generate the image.
- A **discriminator** D evaluates the generated data by its adversary task. In this task, the discriminator is trying to evaluates if the generated image is conditioned on the text caption or not.

In AttnGAN [63], as illustrated in Figure 2.12, 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 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.

The generation and discrimination process is repeated several times in a multistage manner alongside a deep attentional multimodal similarity model

(DAMSM). DAMSM is used to encapsulate that objective in the form of a loss function. Each feature map of the image is assumed to represent the region of the image, where they are treated as equivalent entities with word embeddings from the text caption. Then, the attention mechanism is applied on both the image vector and text vector, in addition to the image feature map with word embeddings. Therefore, the text encoder and image encoder can generate similar embedding, not only in the sentence level but also in the word level.

First, we compute the posterior probability of sentence D_i being matching with image Q_i as follows:

$$P(D_i|Q_i) = \frac{\exp(R_(Q_i, D_i))}{\sum_{j=1}^{M} \exp(R(Q_i, D_j))}, \tag{2.13}$$

where in a batch of training, image Q_i matches with sentence D_i, but the loss is also considering the contrastive condition where the other $M-1$ sentences are a mismatch. Then, the loss function in the word level can be defined as follows:

$$L_1^w = \sum_{i=1}^{M} logP(D_i|Q_i). \tag{2.14}$$

together with the symmetrical definition:

$$L_1^w = \sum_{i=1}^{M} logP(Q_i|D_i). \tag{2.15}$$

After that, we also define the loss function for the sentence level by redefining Equation 2.13 by replacing R with conside similarity to get L_1^s and L_2^s. Therefore, the final DAMSM loss is defined as:

$$L_{DAMSM} = L_1^w + L_2^w + L_1^s + L_2^s \tag{2.16}$$

The generator loss is defined as:

$$L = L_G + L_{DAMSM}, L_G = \sum i = 0m - 1L_{G_i}, \tag{2.17}$$

with m as the number of stages, and L_{G_i} is defined as:

$$L_{G_i} = -\frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[log(D_i(\hat{x}_i))] - \frac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[log(D_i(\hat{x}_i, \bar{e}))], \tag{2.18}$$

where the first part is the unconditional loss to determine if the image is real or fake. The second part is the conditional loss to determine if the image is conditioned to the sentence vector $\bar{e}$ or not.

On the other hand, the loss function for the discriminator is as follows:

$$\begin{aligned} L_{D_i} = &-\tfrac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[log(D_i(x_i))] - \tfrac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[log(1 - D_i(\hat{x}_i))] \\ &-\tfrac{1}{2}\mathbb{E}_{x_i \sim p_{data_i}}[log(D_i(x_i, \bar{e}))] - \tfrac{1}{2}\mathbb{E}_{\hat{x}_i \sim p_{G_i}}[log(1 - D_i(\hat{x}_i, \bar{e}))]. \end{aligned} \tag{2.19}$$

Which is also the combination of unconditional and conditional loss. Each discriminator D_i is trained to classify the input into the class of real or fake by minimizing the loss L_{D_i}. Here, x_i is taken from the image distribution from the data, where $\hat{x}_i$ is taking from the generator hypothesis. This multistage generating and discriminating strategy can successfully synthesize image in detailed clarity that is accurate to the given caption.

Chapter 3

Multimodal Machine Chain (MMC) Framework

This chapter will elaborate on the general definition for the multimodal machine chain framework to address the limited data problem by leveraging feedback from cross-modal mapping (Section 1.2.3). This framework involves various training strategies in different levels of supervision, depending on data availability. Therefore, the multimodal chain framework can maximize the potential of various data conditions.

3.1. Overview of Machine Learning Model Training and Levels of Supervision

A machine learning model has a set of parameters to model the distribution of the data being learned. "Learning" here is achieved by a set of parametric functions that fits the data into the model parameters. A learning process can be regarded as successful if the model can accurately model generalize the training data so that it can predict the label accurately. We define supervision as to how much the training is being guided by the label found inside the data. A model trained with a fully supervised training method requires all the training data to be labelled, while the absence of a label in the data requires the training to be unsupervised.

In this section, we will discuss the difference between levels of supervision in

training a machine learning model.

3.1.1 Supervised Learning

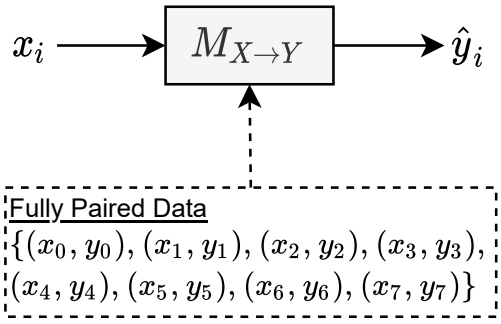

Figure 3.1: Illustration of a model $M_{X \to Y}$ trained with fully-paired data in a supervised manner

In supervised learning, all training data are properly labelled. These labels guide the training process so that the model can change its parameter to try to match the information provided by the feedback [64]. Several tasks that use supervised learning are classification and regression.

Given the definition, supervised model training has less complexity because the labels or number of classes is already known. However, this learning paradigm needs a substantial amount of labelled data, which needs to be collected, sometimes manually. Figure 3.1 illustrates a model training with fully paired data.

3.1.2 Semi-supervised Learning

Semi-supervised learning enables model training from a small amount of labelled data and a large amount of unlabeled data. While the labelled data is being used for training using supervised learning, the unlabeled data is then used to continue the model training. The unlabeled data are usually labelled using the initially trained model, in which this process is called pseudo-labelling. Figure 3.2 illustrates a model training with paired data and unpaired (unlabeled) data.

This learning method enables us to improve a model even when no more paired data are available. Therefore, it is useful for some situations where getting

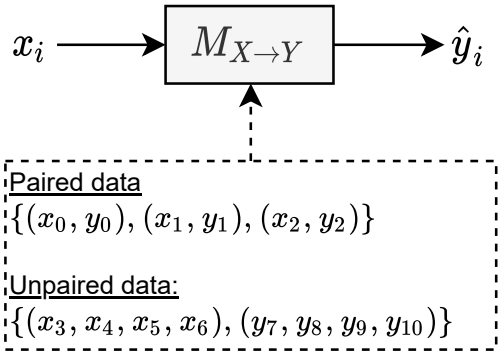

Figure 3.2: Illustration of a model $M_{X \to Y}$ trained with paired data and unpaired data in a semi-supervised manner

a paired data is difficult or costly. However, this method still needs some amount of paired data. In addition, the training steps needs to be differentiated for both labelled and unlabeled data, which increases complexity.

3.1.3 Weakly-supervised Learning

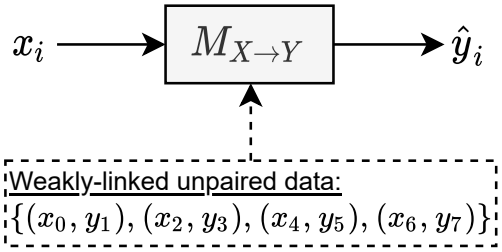

Figure 3.3: Illustration of a model $M_{X \to Y}$ trained with weakly-linked unpaired data in a weakly-supervised manner

This learning strategy significantly reduces the need for paired data, by allowing training with weak supervision. Although in some cases the definition of weakly-supervised learning overlaps with semi-supervised learning, this level of supervision enables learning beyond just using a small amount of paired data. In addition to that, some weak signals or coarse-grained label information can

also be used for learning. Figure 3.3 shows an example of model $M_{X \to Y}$ that is trained with weakly-linked unpaired data. According to Zhou (2017) [65], there are three kinds of weak supervisions covered within this kind of learning:

- **Incomplete supervision:** In this kind of supervision, a small amount of labelled data given is not enough for training the model satisfiably. To further continue the model training, there are two major techniques that can be used: active learning and semi-supervised learning. Semi-supervised learning attempts to use unlabeled data to further improve the model performance altogether with labelled data. This definition has been mentioned in the previous subsection, however, the point here is that semi-supervised learning assumes there is no human intervention contrary to active learning. Active learning assumes there are human experts that can give ground-truth labels as an oracle.

- **Inexact supervision**: This weak supervision defines a situation where the supervision given is not as desired, such as only coarse-grained label information is given.

- **Inaccurate supervision**: This situation is where the supervision given is not always ground-truth. One of the examples of inaccurate supervision is crowdsourcing, where later ground-truth labels are tried to be inferred from the crowd.

Although these three typical weak supervision types are mentioned separately, in practice they may occur simultaneously.

3.1.4 Self-supervised Learning

Self-supervised learning is defined as a process of model learning where the supervision is automatically generated or inferred from the data characteristics itself. This learning style is commonly found in representation learning. For example, in natural language processing, a language model can be trained to predict a missing word from a sentence. In speech processing, a speech representation can be learned by using an autoencoder or denoising autoencoder to reconstruct the

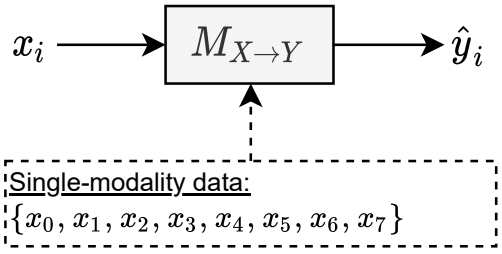

Figure 3.4: Illustration of a model $M_{X \to Y}$ trained with single modality data in a self-supervised manner

speech input. Both examples show that in most cases in self-supervised learning, although the label is found in the data itself, the end product is not the label but the intermediary representation. In the language model example, the end product is how the model assigns a probability to the missing words candidate. On the other hand, the intermediary representation of a speech autoencoder can be used as a better representation for the reconstructed speech. Figure 3.4 illustrates a model that is trained with single-modality data to learn a representation output y.

3.2. General Model Framework

3.2.1 Introduction to Multimodal Machine Chain

In a cross-modal $X \to Y$ mapping task, let the source modality defined 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 $M_{X \to Y}$ based on the data availability. We listed several training strategy, corresponding to each level of supervision related to Section 3.1. Given the data availability, it is also possible to apply several strategies as several steps interchangeably in any order. For example, the training can start from the one with the least supervision to the one with the most supervision, and also its reverses.

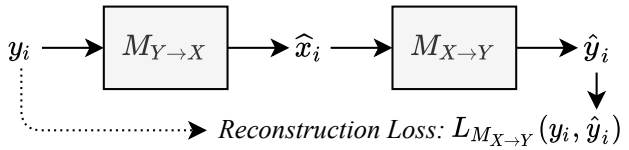

Figure 3.5: Illustration of chain path $\mathcal{C}_{YXY} = \{Y \to X, X \to Y\}$ with $|D| = 2$, where $M_{X \to Y}$ is backpropagated by the reconstruction loss $L_{M_{X \to Y}}$

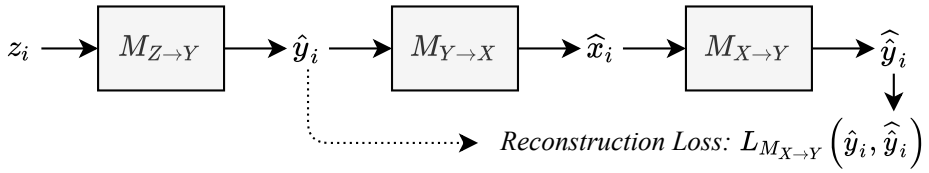

Figure 3.6: Illustration of chain path $\mathcal{C}_{YXY}$ into $\mathcal{C}_{ZYXY} = \{Z \to Y, Y \to X, X \to Y\}$ with $|D| = 3$ for enabling the semi-supervised chain training from single modality data $z_i \in S_z$.

3.2.2 MMC with Fully Paired Data (Supervised Learning)

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

$$\ell_{M_{X \to Y}} = L_{M_{X \to Y}}(y_i^P, \hat{y}_i^P; \theta_{M_{X \to Y}}), \tag{3.1}$$

$$\theta_{M_{X \to Y}} = Optim(\theta_{M_{X \to Y}}, \nabla_{\theta_{M_{X \to Y}}} \ell). \tag{3.2}$$

3.2.3 MMC with Partial Paired Data and Large Amount of Unpaired Data (Semi-supervised Learning)

Since in some cases, there are insufficient data pair in P_{xy} so that $\{(x_0^P, y_0^P), (x_1^P, y_1^P), \ldots, (x_n^P, y_n^P)\} \in P_{xy}, n < m$, cross-modal model $M_{X\to Y}$ cannot be optimally trained to get satisfiable quality. Given unpaired data $\{x_0^U, x_1^U, \ldots, x_m^U\} \in U_x$ and $\{y_0^U, y_1^U, \ldots, y_m^U\} \in U_y$, 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, a chain path $\mathcal{C}_{YXY}$ (See Figure 3.5) can be made to continue training model $M_{X\to Y}$:

$$\mathcal{C}_{YXY} = \{Y \to X, X \to Y\}, \tag{3.3}$$

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

$$\hat{x}_i^U = M_{Y\to X}(y_i^U), \tag{3.4}$$

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

$$\hat{y}_i^U = M_{X\to Y}(\hat{x}_i^U), \tag{3.5}$$

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

$$\ell_{M_{X\to Y}} = L_{M_{X\to Y}}(y_i^U, \hat{y}_i^U; \theta_{M_{X\to Y}}). \tag{3.6}$$

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

$$\theta_{M_{X\to Y}} = Optim(\theta_{M_{X\to Y}, M_{Y\to X}}, \nabla_{\theta_{M_{X\to Y}, M_{Y\to X}}}\ell), \tag{3.7}$$

while in a non end-to-end condition, Eq. 3.2 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.

3.2.4 MMC with Partial Paired Data, Few Amount of Unpaired Data, and Unrelated Single Modality Data (Semi-supervised Learning)

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 Section 3.2.3. 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 $\{(x_0^P, y_0^P, z_0^P), (x_1^P, y_1^P, z_1^P), \ldots, (x_n^P, y_n^P, z_n^P)\} \in P_{xyz}, n < m$, which are inadequate to satisfiably train $M_{X\to Y}$ as in Section 3.2.2. Similar to Section 3.2.3, single-modality data $\{x_0^S, x_1^S, \ldots, x_m^S\} \in S_x$, $\{y_0^S, y_1^S, \ldots, y_m^S\} \in S_y$, and $\{z_0^S, z_1^S, \ldots, z_m^S\} \in S_z$ are available. In this condition, a chain path C_{ZYXY} that leverages S_Z single-modality data can be constructed as follows:

$$\mathcal{C}_{ZYXY} = \{Z \to Y, Y \to X, X \to Y\}, \tag{3.8}$$

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

$$\hat{y}_i^S = M_{Z\to Y}(z_i^S), \tag{3.9}$$

$$\hat{x}_i^S = M_{Y\to X}(\hat{y}_i^S), \tag{3.10}$$

$$\hat{\hat{y}}_i^S = M_{X\to Y}(\hat{x}_i^S), \tag{3.11}$$

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

$$\ell_{M_{X\to Y}} = L_{M_{X\to Y}}(\hat{y}_i^S, \hat{\hat{y}}_i^S; \theta_{M_{X\to Y}}). \tag{3.12}$$

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

$$\theta_{M_{X\to Y}} = Optim(\theta_{M_{X\to Y}, M_{Y\to X}, M_{Z\to Y}}, \nabla_{\theta_{M_{X\to Y}, M_{Y\to X}, M_{Z\to Y}}} \ell), \tag{3.13}$$

while in a non end-to-end condition, Eq. 3.2 is sufficient. Figure 3.6 illustrates this chain path.

As we can see from the process flow, Eq. 3.10-3.13 are similar with Eq. 3.4-

3.7 because the chain path $\mathcal{C}_{YXY}$ are inside the path of $\mathcal{C}_{ZYXY}$. Therefore, an extension from the chain with $|D| = 2$ to $|D| = 3$ can be developed, which further shows the generalization of the chain framework.

3.2.5 MMC with Fully Unpaired Data (Weakly-supervised Learning)

In a case where there are no paired data P_{xy} available, a cross-modal model $M_{X\to Y}$ can still be trained with unpaired data $U_{x,y}$ if there is some weak supervision available to connect X to Y. The weak supervision here are implemented as a pivot that provides bridging information $B_{X\to Y}$. $M_{X\to Y}$ training will be as follows:

$$\hat{y}_i^U, \hat{y}_i^{alg} = M_{X\to Y}(x_i^U, B_{X\to Y}), \tag{3.14}$$

where model $M_{X\to Y}$ generates both the $\hat{y}_i^U$ hypothesis and $\hat{y}_i^{alg}$ alignment hypothesis. Both hypotheses can be used to calculate the supervised loss:

$$\ell_{M_{X\to Y}}^{sup} = L_{M_{X\to Y}}(y_i^P, \hat{y}_i^P; \theta_{M_{X\to Y}}), \tag{3.15}$$

in addition to the alignment loss:

$$\ell_{M_{X\to Y}}^{alg} = L_{M_{X\to Y}}(y_i^{alg}, \hat{y}_i^{alg}; \theta_{M_{X\to Y}}). \tag{3.16}$$

Then, by summing both of the loss $\ell_{M_{X\to Y}} = \ell_{M_{X\to Y}}^{sup} + \ell_{M_{X\to Y}}^{alg}$, model $M_{X\to Y}$ can be backpropagated as follows:

$$\theta_{M_{X\to Y}} = Optim(\theta_{M_{X\to Y}}, \nabla_{\theta_{M_{X\to Y}}}\ell). \tag{3.17}$$

3.2.6 MMC with Only Single Modality Data (Self-supervised Learning)

On the other hand, there might exist a data where a usable representation is not available. In the previous example, the modality X does not have a proper representation to allow mapping to be done effectively. In this case, a new representation of X^D can be learned using a self-supervised learning strategy with

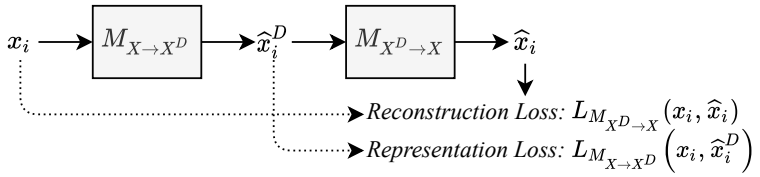

Figure 3.7: Illustration of chain path $\mathcal{C}_{XX^DX} = \{X \to X^D, X^D \to X\}$ with $|D|$ =2 to learn new representation $\hat{x}_i^D \in X^D$ using reconstruction loss $L_{M_{X^D \to X}}$ and representation loss $L_{M_{X \to X^D}}$

using a single modality data $\{x_0^S, x_1^S, \ldots, x_m^S\} \in S_x$. Let $\{\hat{x}_0^D, \hat{x}_1^D, \ldots, \hat{x}_m^D\} \in D_x$ be defined as the new representation of X. A chain path $\mathcal{C}_{XX^DX}$ (See Figure 3.7) can be made:

$$\mathcal{C}_{XX^DX} = \{X \to X^D, X^D \to X\}, \tag{3.18}$$

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

$$\hat{x}_i^D = M_{X \to X^D}(x_i^S), \tag{3.19}$$

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

$$\hat{x}_i^S = M_{X^D \to X}(\hat{x}_i^D), \tag{3.20}$$

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

$$\ell_{M_{X \to X}} = L_{M_{X \to X^D} M_{X^D \to X}}(x_i^S, \hat{x}_i^S; \theta_{M_{X \to X^D}, M_{X^D \to X}}), \tag{3.21}$$

in addition to the representation loss:

$$\ell_{M_{X \to X^D}} = L_{M_{X \to X^D}}(x_i^S, \hat{x}_i^D; \theta_{M_{X \to X^D}}). \tag{3.22}$$

This training strategy only needs single-modality data such as S_x, which can be classified as self-supervised training.

Chapter 4

MMC Framework for Cross-modal Collaboration through Listening, Speaking, and Visualizing

This chapter describes the use of the MMC framework for a semi-supervised cross-modal collaboration (MMC-SemiSup) in between several cross-modal models. First, each of the cross-modal models is independently trained using supervised learning as described in Section 3.2.2 using a small amount of paired data. Second, we take the advantage of the available unpaired data to train the chain with the framework learning strategy as mentioned in Section 3.2.3. Finally, we show our general framework capability to enable semi-supervised training using single-modality data from unrelated modality (Section 3.2.4).

4.1. Introduction

The machine speech chain [50] was successfully enabling ASR and TTS model training from an unpaired dataset. However, unlike human communication which is multimodal, it is still unclear how to incorporate other modalities such as visual modalities in the chain. In addition, the modalities of the unpaired data being

used are the same modalities of the input and output (i.e. speech and text). Therefore, the chain mechanism itself is still limited to the task, as described in Section 1.2.3.

In this chapter, we developed the generalization of this speech chain mechanism into a semi-supervised chain that can generate feedback from any modalities. Every time an input is converted to another modality and converted back again, it generates feedback in the form of reconstruction loss. This feedback mechanism is inspired by human communication that does not need parallel data, as described in Section 1.2.3. Therefore, a cross-modal model training can be continued with single-modality data from a modality that is even unrelated to the cross-modal task itself (i.e. image data to train ASR model). We use the definition of multimodal machine chain framework in Chapter 3, with ASR, TTS, IC, IR and IG model described in Chapter 2.

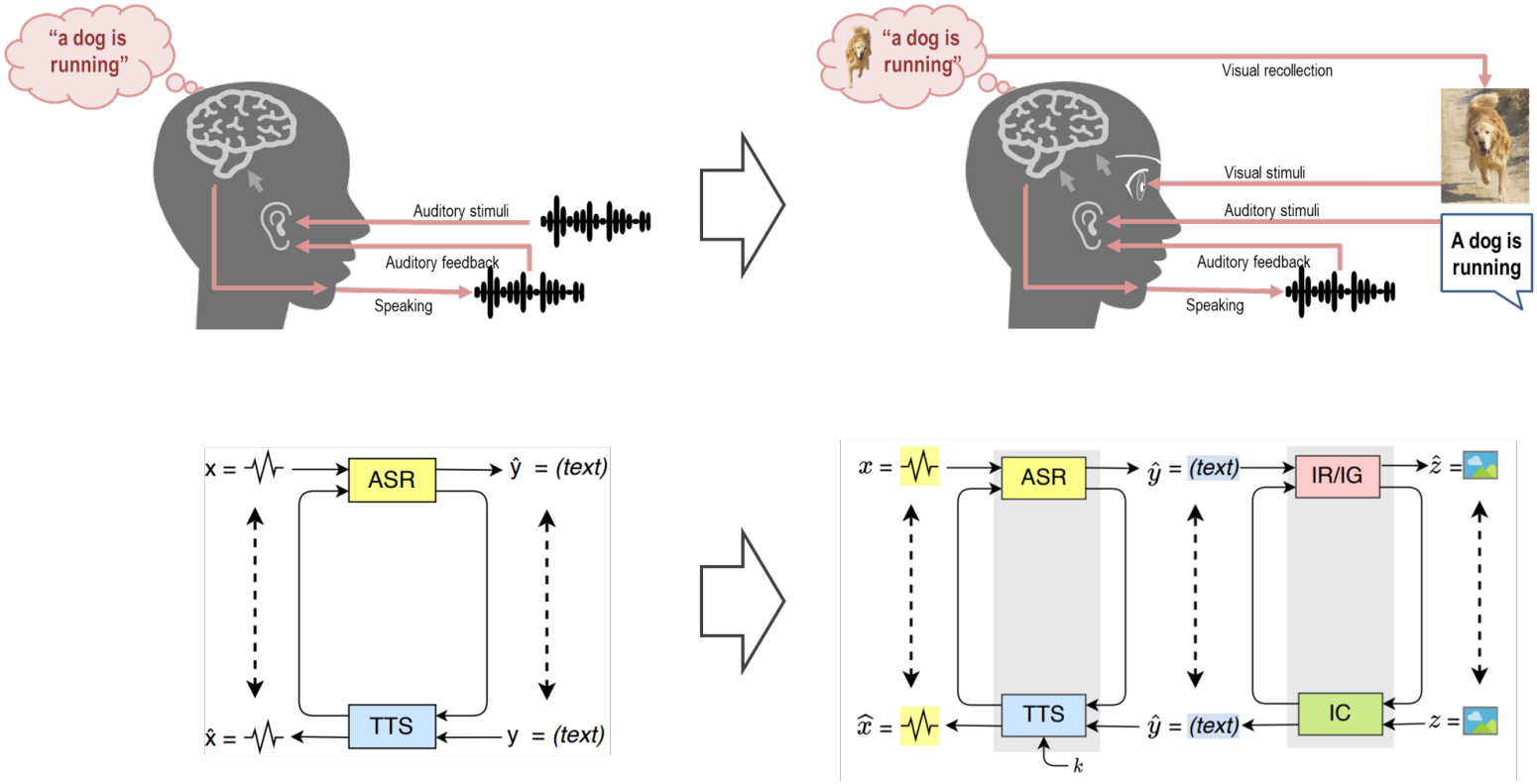

Figure 4.1: Generalizing the chain mechanism: from speech chain to multimodal chain for semi-supervised cross-modal collaboration (MMC-SemiSup)

4.2. Previous Work

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. [37]. A popular extension of the LAS framework [21] (Section 2.1.1), called "Watch, Listen, Attend, and Spell (WLAS)," was proposed by Chung et al. [66]. 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. [38] also proposed a deep audio-visual speech recognition system to recognize phrases and sentences from a talking face. 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. [67] 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, which also briefly mentioned in Section 1.2.3 as the curse of dimensionality, 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, as mentioned in Section 1.2.3. The speech chain framework [50, 51, 52, 53] might be the first framework constructed on different modality domains (speech versus text). Then, Karita et al. (2019), proposed a semi-supervised ASR and TTS, that are using autoencoders to enable joint representation [68]. In the image-to-text domain, Turbo Learning combined image

captioning and generation in a joint training framework [69]. In this chapter, we propose the use of the MMC framework to accommodate the triangle modality and the loop feedback mechanism.

4.3. Semi-supervised Multimodal Chain Framework Cross-modal Collaboration (MMC-SemiSup)

4.3.1 Previous Work: Machine Speech Chain

This section describes the machine speech chain [50, 51, 52, 53] 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 3.2 to describe the basic 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.

4.3.2 Proposed: Dual-loop MMC-SemiSup

The generalization of a semi-supervised chain mechanism when $|D| = 3$ is realized with three kinds of modalities, X, Y, Z for speech, texts, and images. To connect each of these modalities in this MMC-SemiSup, we define five kinds of models and our proposed chain path to improve them in a semi-supervised manner with single-modality data:

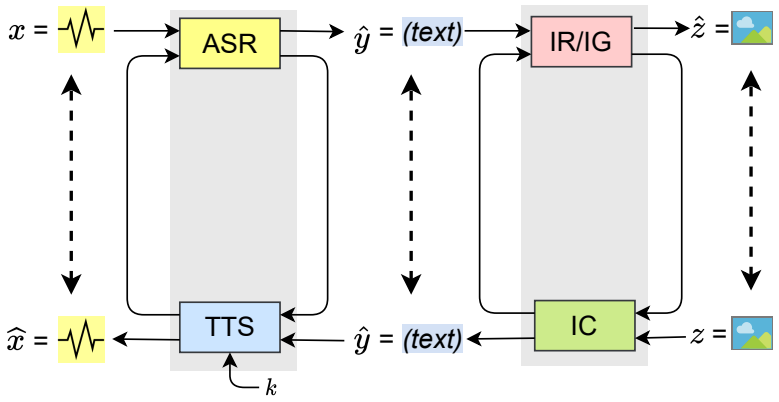

Figure 4.2: Dual-loop multimodal chain for cross-modal collaboration with image retrieval (IR) or image generation (IG) (MMC-SemiSup1-IR/IG)

- $M_{X \to Y}$ 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),
- $M_{Z \to Y}$ 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 4.2(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. This multimodal chain for cross-modal collaboration is called **MMC-SemiSup1-IR**, when the visual chain is using an IR model, and **MMC-SemiSup1-IG**, when the visual chain is using an IG model.

The training steps for MMC-SemiSup1-IR and MMC-SemiSup1-IG are as follows:

- **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. (Section 3.2.2)

- **Step 2: semi-supervised training using unpaired data**
  The training can be continued in a semi-supervised manner using unpaired

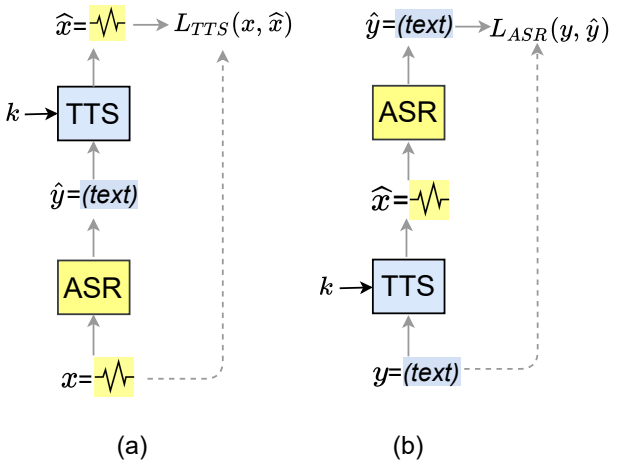

Figure 4.3: Unrolled process for speech chain, when the input is (a) speech or (b) text.

image-speech-text data $U_{x,y,z}$, as described in Section 3.2.3. In the speech chain (Figure 4.3), $\mathcal{C}_{YXY}$ 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 (Figure 4.4), $\mathcal{C}_{ZYZ}$ 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**
  Using the learning mechanism described in Section 3.2.4, given speech only data S_x, two chain paths can be made: $\mathcal{C}_{XYX}$ 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

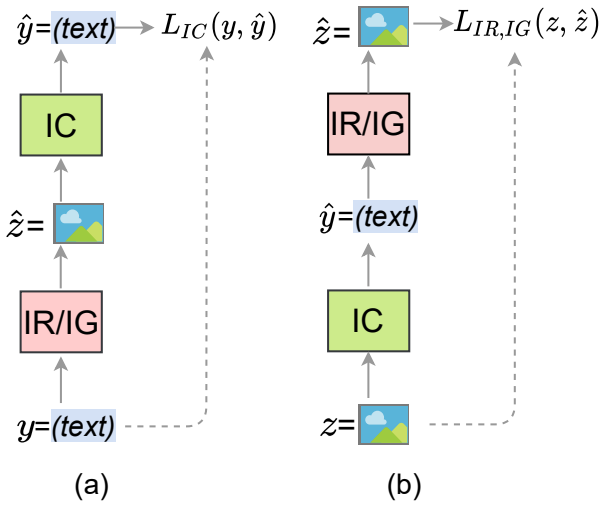

Figure 4.4: Unrolled process for visual chain, when the input is (a) text or (b) image.

$M_{Y \to Z}$ 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 $\mathcal{C}_{ZYXY}$ (Figure 4.5). 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. The 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.

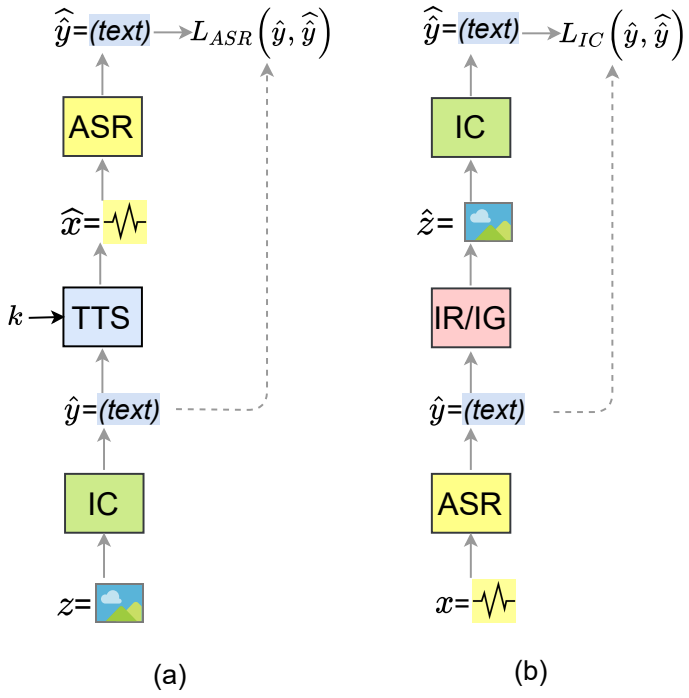

Figure 4.5: Unrolled process for cross-modal collaboration between speech and visual chain (MMC-SemiSup1), when the input is (a) image or (b) speech.

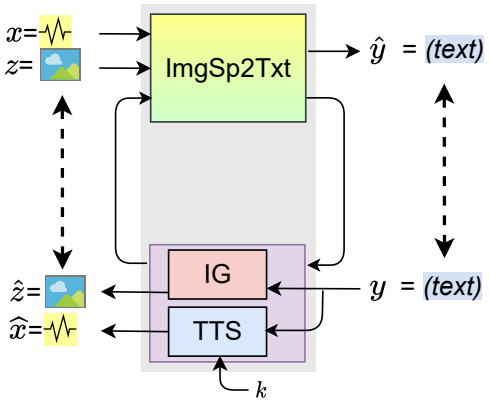

Figure 4.6: Single-loop MMC-SemiSup for cross-modal collaboration (MMC-SemiSup2)

4.3.3 Single-loop MMC-SemiSup

Next, we propose a single-loop multimodal chain for cross-modal collaboration (MMC-SemiSup2) to show the implementation of our proposed chain framework

in a multi-source multimodal model environment. In this kind of MMC-SemiSup, ASR and IC model is combined to promote sharing between these two models. Therefore, the loop mechanism resembles a chain implementation when $|D| = 2$ (Section 3.2.3), 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 4.6, 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 input.

- **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 MMC-SemiSup2 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, two chain paths can be defined: $\mathcal{C}_{GYG}$ and $\mathcal{C}_{YGY}$, resembling chain paths when $|D| = 2$.

  The first path $\mathcal{C}_{GYG}$ is used when MMC-SemiSup2 is given either unpaired image-speech-text dataset $U_{x,y,z}$, speech-only dataset S_x, or image-only dataset S_z (Figure 4.7). 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 backpropagated by

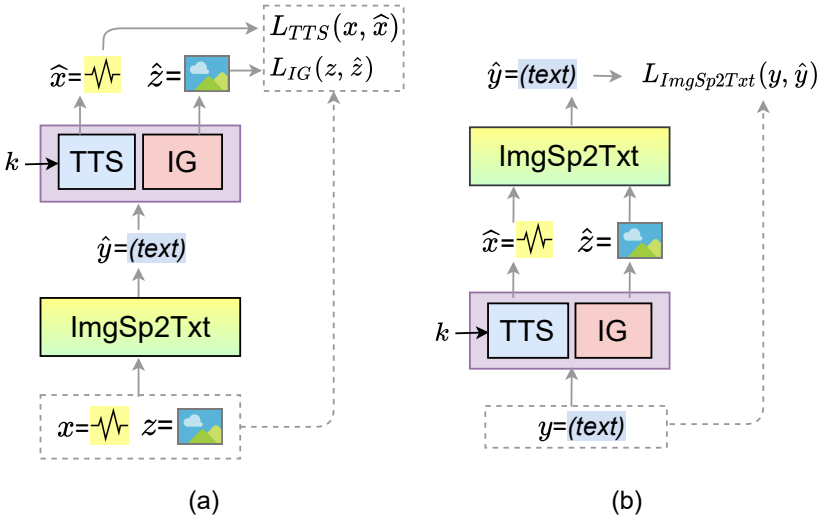

Figure 4.7: Unrolled process for single-loop MMC-SemiSup, when the input is (a) speech/image or (b) text.

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 $M_{G \to Y}$ ImgSp2Txt model using the reconstruction loss $L_{M_{G \to Y}}$ by comparing the original text y and text hypothesis $\hat{y}$.

4.3.4 MMC-SemiSup Components

- **ASR**
  We use the ASR model described in Section 2.1.1. The input is 80 dimensional mel-spectrogram, while the output is text transcription with character level granularity. The attention mechanism is using multilayer perceptron (MLP) attention.

- **TTS**
  The TTS model is described in Section 2.1.2. Instead of the phoneme, we

use the character as input to math the output feature of ASR, similar to Tjandra et al. [50]. The output of the sequence-to-sequence model is an 80 dimensional mel-spectrogram and a stop token.

- **IC**
  IC model for this chapter is build resembling the LSTM-based Show, Attend, and Tell model, as described in Section 2.2.1.

- **IR**
  We train a shared embedding between image and text for image retrieval, as described in Section 2.2.2.

- **IG**
  We use AttnGAN model described in Section 2.2.3 for image generation model.

- **Two-fold image-speech to text**
  MMC-SemiSup2 is designed to show the effectiveness of the cross-modal collaboration mechanism in a multi-source multimodal model environment. We combine ASR and IC model to create a model that receives image and speech information and generates the text transcription/caption, as the example of a multi-source multimodal model which we call ImgSp2Txt.

  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 4.8). In this model, the image is encoded by a residual network that produces high-level feature representation $\mathbf{e^z} = [e_0^z, .., e_n^z]$ of the image. Bidirectional LSTM encodes the speech features into embedded representation

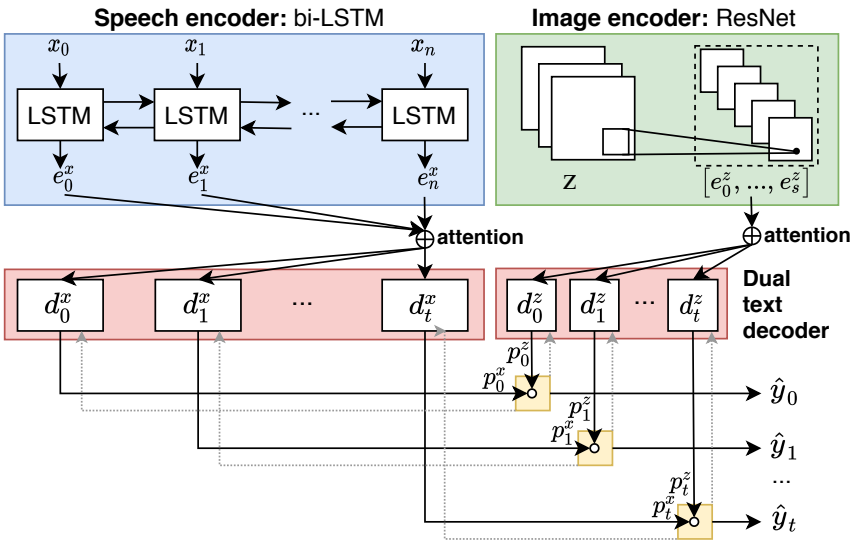

Figure 4.8: Dual text decoder with audio and visual decoding combination

$\mathbf{e^x} = [e_0^x, .., e_m^x]$. 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.

4.4. Experiment Settings

4.4.1 Dataset

1. Flickr 30k
   Flickr30k [70] is an image-captioning dataset which images from Flickr consist of everyday activities, scenes, and events. There are about 150k crowdsourced 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. The Google TTS is just used to generate the dataset, and not for training.

2. Flickr 8k

Similar to Flickr30k, Flickr8k [71] 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 crowdsourcing platform. This dataset has 183 unique speakers.

4.4.2 Dataset Composition

Table 4.1: 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)

The default dataset split is used for Flickr30k (29k train, 1k dev, and 1k test) and Flickr8k (6k train, 1k dev, and 1k test). A scenario is designed to showcase the multimodal machine chain ability to improve model quality in a semi-supervised manner using a single modality dataset. For Flickr8k, all the five captions from an image are used, while in Flickr30k, the same settings as the previously published work [72] are used to balance the image production side.

Table 4.1 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 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 4.2. Before partitioning the data, we randomly shuffle the order of the keys in the dataset initially. For measuring the topline performance such as in Section 4.5.1, 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 cross-modal collaboration, 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 from Section 3.2.2, Section 3.2.3, and Section 3.2.4. With these trained models, we can compare which image production method is better for the MMC-SemiSup: IR or IG. As listed in Table 4.2, the Flickr30k dataset contains 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 MMC-SemiSup can also work in a multi-speaker natural speech dataset. We tested MMC-SemiSup2 with the same data partition to compare it with a label propagation method (Section 4.4.5). 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 3.2.4). The remaining data (other than the paired data) were regarded as a 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, 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 cross-modal collaboration 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.

Table 4.2: Data partitioning for each subset (in #Image (hours)). $n = \{0, 1, 2, 3, 4, 5\}$, $m = \{0, 1, 2, 3, 4, 5, 6, 7\}$

Task	**Note**	**Section**	**Paired**	**Unpaired**	**Single-modality**		**Total**
			$\#P_{xyz}$ (hours)	$\#U_{x,y,z}$ (hours)	$\#S_x$ (hours)	$\#S_z$ (hours)	(hours)
Dataset: Flickr30k							
Topline	For comparison with existing published systems	4.5.1	29000 (51.96)	0	0	0	29000 (51.96)
IR vs IG	For comparison between MMC-SemiSup1-IR and MMC-SemiSup1-IG	4.5.2	2000 (3.54)	7000 (12.55)	10000 (19.97)	10000 (17.89)	29000 (51.96)
Dataset: Flickr8k							
Topline	For comparison with existing published systems	4.5.1	6000 (34.31)	0	0	0	6000 (34.31)
(Label Prop. I) & (MMC-SemiSup1 vs MMC-SemiSup2)	For comparison between label propagation and our proposed MMC-SemiSup	4.5.3&4.5.4	800 (4.57)	1500 (8.57)	1850 (10.70)	1850 (10.56)	6000 (10.56)
Label Prop. II	For comparison with label propagation using more paired data	4.5.3	1400 (8.00)	900 (3.43)	1850 (10.70)	1850 (10.56)	6000 (34.31)
No single modality	For checking performances when all remaining data other than paired are unpaired	4.5.4	800 (4.57)	5200 (29.74)	0	0	6000 (34.31)
Var. single modality	For investigation of effect of increasing single-modality data amount	4.5.5	800 (4.57)	0	$520n$ (5.95n)	$520n$ (5.95n)	800-6000 (4.57-34.31)
Var. paired	For investigating the effect of increasing initial paired data amount	4.5.6	$200+300m$ (1.14+1.86m)	0	1850 (10.70)	1850 (10.56)	3900-6000 (13.14-34.31)

4.4.3 Model Details

We implemented all the models described in Section 4.3.4.

Speech Processing Models

We used an 80-dimensional mel-spectrogram for the speech features in the ASR, TTS, and ImgSp2Txt models. We used character-level granularity for ASR, TTS, and ImgSp2Txt tasks. The ASR model is a standard Listen, Attend, and Spell model with multi-layer perceptron (linear) location-aware attention [21]. The encoder part consists of bidirectional LSTM with the depth of 3 and the size of 256 for one direction. The forward and backward LSTM is concatenated to get the final encoder hidden representation. We use pyramidal mechanism in between layers, to reduce the length of the encoded speech feature. With the depth of three, one encoder hidden representation represents 2^3 number of frames. The decoder is an LSTM decoder with 512 size and depth 1. Both encoder and decoder has a dropout probability of 0.25. We use label smoothing with a value of 0.05 for the output layer. We trained the ASR model using Adam [73] with 1e-3 learning rate. The ImgSp2Txt model uses the same parameter as ASR, with a combination with IC in the output layer. Both ASR and IC has the same probability (50:50).

Moreover, we used the TTS model described in Section 2.1.2, which is similar to the Tacotron [2]. We use our own Tacotron implementation and trained it from the initial state (i.e. not pretrained). We added a speaker embedding with a dimension of 64 as the condition for the decoder and the start-stop prediction. The encoder is a CBHG encoder with the prenet size of 256 and dropout probability of 0.5. Then, we use historical attention [74], similar with the one used in [51]. To handle unseen speakers when processing a non-paired data, we performed one-shot speaker adaptation from the pool of paired speech data. The decoder uses LSTM layer with depth one and hidden size of 512. The stop prediction module use a hidden size of 256 and the output layer of one. Finally, to generate a waveform, we use the inverter part of the Tacotron with the projection size of 256 to generate a linear-spectrogram from the generated mel-spectrogram. After that, we use Griffin-Lim algorithm to finish the waveform generation process, as described in Section 2.1.2. We trained the TTS model using Adam [73] with

2.5e-4 learning rate.

Image Processing Models

We implemented the IC model described in Section 2.2.1, which resembles the Show, Attend, and Tell model [61]. We used a pretrained Resnet-50 [4] on ImageNet task [75] as the image encoder, by removing the last two layers to get a high-level feature with the size of 512. Then, using multilayer perceptron attention with the hidden size of 512, we decoded the caption using LSTM decoder with the depth of one, hidden size of 512, and dropout probability of 0.5. We trained the IC model with Adam [73] optimizer, with the learning rate of 1e-4.

The IR model uses a shared embedding between image and text, with a dimension of 300. The text embedding is generated using an LSTM encoder, with a depth of one and hidden size of 512. Then, we use an image encoder with similar specification of IC, to get the image embedding. We removed the last layer of ResNet [4], to get a single image vector with the size of 2048. Both image and text vector are then projected in to a 300 dimension using a linear layer. We train IR model with stochastic gradient descent with a 0.1 learning rate.

The IG model resembles AttnGAN [63], as described in Section 2.2.3. We use the same parameters as in their paper. We reduces the number of step into two steps from three steps, which consequently reduces the image size into 128x128 so that it can reduces memory consumption. The DAMSM mechanism is only trained on the paired dataset, with a class size of 50. We trained the IG model using Adam optimizer with the learning rate of 2e-4 for both the generator and the discriminator.

Chain Mechanism

Although technically training each element in the chain path is possible with an end-to-end style [52], 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. When the training failed to reduce the development loss after a warmup of five, we halved the learning rate. The training stops when the development loss does not decreased anymore, with the average of 50 epochs.

4.4.4 Evaluation Metrics

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) [76] 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) [77] to determine how realistic the IG output was.

4.4.5 Label Propagation

Label propagation [44] 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 [45]. 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.

4.5. Experiment Result and Analysis

4.5.1 Topline Scenario

In this section, we simulated a condition where much paired data are available (Table 4.2: Topline). Our experiment measured our model performance and compared it to previously published studies. In addition, we also listed the method for each model reported. 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. The purpose of this comparison is to confirm whether the model used in our framework performs as well as the on reported on the previously publised studies.

We listed all the scores of our topline model in Table 4.3. In the WSJ corpus [78], our ASR model performed as well as Kim et al.'s result using JointCTC+Attention [74]. In addition, we also found that our ASR model can perform as good as Bahdanau et al.'s ASR model, which has a similar architecture of encoder-decoder with attention. Then, ASR and TTS work as well as the previously published results of Tjandra et al. [80, 52]. Our IC model performed as good as Xu et al. [61] 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. For image processing related model, we use image augmentation strategies which increases the model performance.

4.5.2 Proposed: From IR to IG

First, we need to decide whether the IR or the IG model is better for the MMC-SemiSup. 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.

Table 4.3: Comparison of our model performances with existing published results: ↓ means lower is better; ↑ means higher is better.

Data	Model	Method	Result
ASR - CER (%) ↓			
WSJ [78]	Kim et al. [74]	Content-based Att	11.08
	Kim et al. [74]	JointCTC + Att	7.36
	Bahdanau et al. [79]	EncDec + Att	6.4
	Tjandra et al. [80, 52]	EncDec + Att	6.43
	Ours (Sec. 2.1.1)	EncDec + Att	6.60
TTS - L2 ↓			
WSJ	Tjandra et al. [80]	Tacotron	0.64
	Ours (Sec. 2.1.2)	Tacotron	0.68
IC - B1/B4 ↑			
Flickr8k	Xu et al. [61]	SAT	66.90 / 19.90
	Ours (Sec. 2.2.1)	SAT + augment	65.93 / 22.56
IR - R@10 ↑			
Flickr30k	Vilalta et al. [81]	emb-based IR	59.8
	Ours (Sec. 2.2.2)	emb-based IR + augment	62.42
IG - Inception ↑			
CUB	Xu et al. [63]	AttnGAN	4.36
	Ours (Sec. 2.2.3)	AttnGAN + augment	5.67
ImgSp2Txt - CER / WER (%) ↓			
Flickr8k	Sun et al. [67]	Image-LM	- / 13.81
	Ours (Sec. 4.3.4)	Img+Sp Ensemble	5.16 / 7.13

Table 4.4: Comparison of performance of proposed MMC-SemiSup1-IR with MMC-SemiSup1-IG on Flickr30k

Training	Data Type	#Image	ASR CER↓	IC B1/B4↑	TTS L2[2]↓	IR R@10↑	IG IS↑
MMC-SemiSup1-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	-
MMC-SemiSup1-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

For this, we used MMC-SemiSup1 and replaced the image production model using IR or IG. We labelled each of them as MMC-SemiSup1-IR and MMC-SemiSup1-IG. We partitioned the data for Steps 1, 2, and 3 following the steps in

Section 4.3. For the size of each subset, refer to Table 4.2:"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 4.4, 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 MMC-SemiSup1-IR improvement in Step 2 is more focused on B1 than B4, compared with MMC-SemiSup1-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 MMC-SemiSup1-IG outperformed MMC-SemiSup1-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 MMC-SemiSup1-IR shows a performance decrease, although in MMC-SemiSup1-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 MMC-SemiSup1-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.

4.5.3 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 4.5. 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 MMC-SemiSup, 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 MMC-SemiSup is more effective than the label propagation method.

Table 4.5: Comparison of proposed MMC-SemiSup1 and MMC-SemiSup2 performances with label propagation method in Flickr8k dataset

Training	**Data Type**	**#Image**	MMC-SemiSup1-IG				MMC-SemiSup2			
			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
Cross-modal Collaboration	$+U_{x,y,z}$ 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
Cross-modal Collaboration	$+U_{x,y,z}$ 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

4.5.4 Proposed: Comparing MMC-SemiSup1-IG and MMC-SemiSup2

After choosing between IR and IG and comparing with the label propagation baseline, in this section we evaluate the performance of a dual-loop (MMC-SemiSup1-IG) vs. a single-loop (MMC-SemiSup2) multimodal chain for cross-modal collaboration. For the data partitioning in this experiment, we refer to the subset partitioning based in Table 4.2: MMC-SemiSup1 vs MMC-SemiSup2. Initially, we separately trained all the models using P_{xyz} data in a supervised manner. As shown in Table 4.5, both MMC-SemiSup1-IG and MMC-SemiSup2 have 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 MMC-SemiSup1-IG and MMC-SemiSup2 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 4.5, 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 MMC-SemiSup1 or in a single-loop MMC-SemiSup2.

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 MMC-SemiSup. For MMC-SemiSup1-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 a single modality (See Table 4.5: Separated(Semi-supervised)). In a single-loop MMC-SemiSup2, however, our proposed method remains superior. With our proposed cross-modal collaboration, 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 [67]. Although our proposed approach is semi-supervised, we can still achieve a comparable error rate to a fully-supervised ASR system.

4.5.5 Single modality data amount effect to the final speech processing model performance

Our proposed multimodal chain for cross-modal collaboration 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 improve consistently as the amount of single modality data increases. For this additional experiment, we refer to the data partitioning shown in Table 4.2: Var. Single Modality.

Figure 4.9 compares the ASR improvement using MMC-SemiSups 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 MMC-SemiSup. The best CER score (23.07%) was reached using all of the single modality data of 2600 images.

In addition, Figure 4.10 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 the cross-modal collaboration is positively related to how many

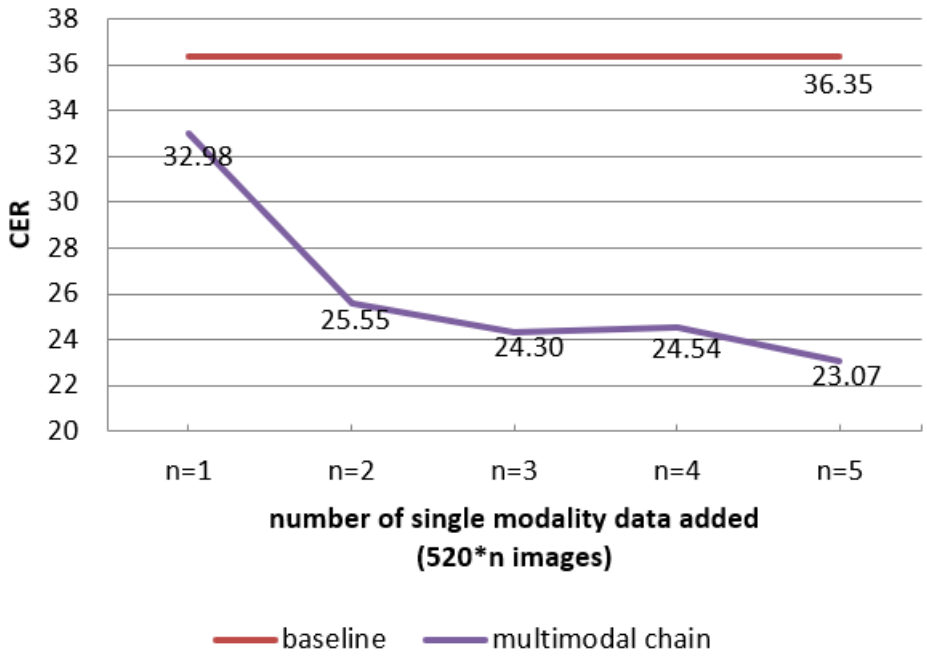

Figure 4.9: Single modality data amount effect to final ASR performance compared with initial model baseline in Flickr8k natural speech dataset. Vertical axis: character error rate (CER). Horizontal axis: number of single modality data added.

more data are used in the semi-supervised step by leveraging the cross-modal augmentation.

4.5.6 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 4.2: 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 the cross-modal collaboration to improve the performance in semi-supervised steps, we measured the score differences between the initial model and the model after the semi-supervised step.

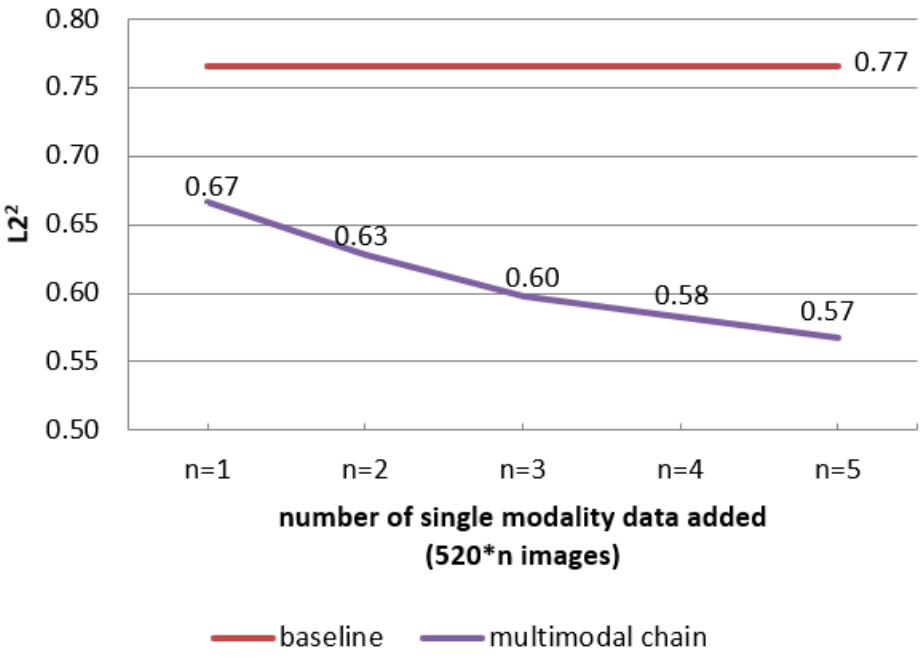

Figure 4.10: Single modality data amount effect to final TTS performance compared with initial model baseline in Flickr8k natural speech dataset. Vertical axis: $L2^2$ Loss. Horizontal axis: number of single modality data added.

Table 4.6: 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 *MMC-SemiSup*	$\triangle$CER $\uparrow$ *MMC-SemiSup*
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

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 semi-supervised step. In this scenario, a larger amount of initial data

Table 4.7: 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 *MMC-SemiSup*	$\triangle$L2^2 $\uparrow$ *MMC-SemiSup*
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

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 cross-modal collaboration started to show its effectiveness in improving the ASR model performance, manifested by positive $\triangle$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 $\triangle$L2^2 score remained positive when the initial data sizes exceed 500, suggesting that our proposed cross-modal collaboration improved the TTS performance even when the initial TTS model was already 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. With these results, we can conclude that the minimum paired data needed for convergence in Flickr8k is about 4000 utterances (about 4.57 hours). This is because such an amount of data will enable ASR training with the accuracy of about 40% CER in the Flickr8k multispeaker natural speech dataset (Table 4.6). In addition, we can also refer to Tjandra et al.'s machine speech chain [54], in Table 1, where they reported that some improvement is still possible even with 2 hours of single-speaker data (LJSpeech). This is because it enables a baseline with a 31.7% initial ASR model performance using only 10% of the total data (about 1200 utterances). Therefore, we may conclude that the baseline model performance

shall achieve about a 40% error rate or less.

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 cross-modal collaboration 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.

4.6. Summary

In this chapter, 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 proved our argument in Section 1.2.3 that with our proposed framework, adding modality will enable more feedback for training, instead of increasing training difficulties due to limited parallel data problem. 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 MMC-SemiSup in a multispeaker natural speech dataset. Our chain mechanism can be implemented on an audiovisual model through a single-loop

MMC-SemiSup, 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 cross-modal collaboration 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 cross-modal collaboration is more effective in a low-resource scenario, when the initial paired data are insufficient to satisfiably train the cross-modal model

Chapter 5

MMC Framework for Speech-to-text Mapping using Visually-connected Non-parallel data

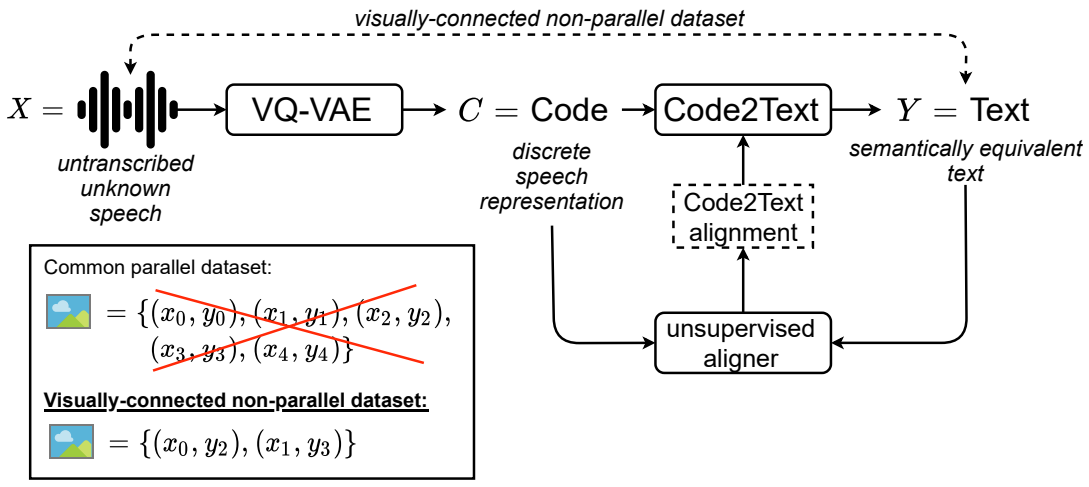

Figure 5.1: Multimodal machine chain framework for weakly-supervised speech-to-text mapping (MMC-WeakSup)

The previous chapter has described our attempt in generalizing the machine

speech chain into any kind of modalities and any kind of data availability, in the form of MMC-SemiSup. However, such a chain mechanism still needs paired data to initialize the cross-modal model. In this chapter, we attempt to use the MMC framework for speech-to-text mapping using visually-connected non-parallel data using part of the general framework mentioned in Section 3.2.5. We call this "mapping" since the system attempts to learn the semantic association between speech and text instead of recognizing the speech with the exact word-by-word transcription. Unlike MMC-SemiSup, this MMC application does not need paired data. Referring to the levels of supervision described in Section 3.1, our proposed approach in this chapter can be categorized as a weakly-supervised learning with inexact supervision. We call this application of MMC as **MMC-WeakSup**, where our proposed model learns to map speech-to-text by exploiting the partial overlap between the two, based on the fact that both of the speech and text are both visually-connected.

MMC-WeakSup is implemented as a novel cyclic partially-aligned Transformer with two-fold mechanisms (See Figure 5.1). First, we train a Transformer-based vector-quantized variational autoencoder (VQ-VAE) to produce a discrete speech representation in a self-supervised manner. Then, we use a Transformer-based sequence-to-sequence model inside a chain mechanism to map from unknown untranscribed speech utterances into a semantically equivalent text. Because this is not strictly recognizing speech, we focus on evaluating the semantic equivalence of the generated text hypothesis. Our evaluation shows that our proposed method is also effective for a multispeaker natural speech dataset and can also be applied for a cross-lingual application.

5.1. Introduction

In human communication, it often does not matter whether we can figure out word-by-word what the speaker is saying as long as we understand the semantic message the speaker wants to convey. Therefore, we argue that it may be possible to address the construction of spoken language processing without having speech utterances and the exact corresponding transcriptions, which are generally unavailable. In fact, there are many available collections of texts and pictures from online books, and there are many available speeches recorded with images/videos in social media (i.e., YouTube). If we could link to those images, we might be able to create visually connected non-parallel speech-text data.

This study addresses weakly-supervised speech-to-text mapping problem given only a collection of visually connected non-parallel speech-text data. This may be considered one of the new ways of building speech-to-text transformation systems within a language but without using ASR. The system learns the semantic association between speech and text instead of recognizing the content of speech utterances with an exact word-by-word transcription. It can also be considered as a paraphrasing or translation task from unknown untranscribed speech utterances into semantically equivalent texts. Since this system does not strictly recognize speech, we focus on evaluating the semantic equivalence of the generated text hypothesis.

5.2. Related Work

As mentioned in Section 1.2.3, research on constructing technologies with less or without parallel data has been gained attention. To date, various approaches have been proposed for developing voice conversion systems with non-parallel data [82, 83, 84, 85]. One approach applies unsupervised neural machine translation to develop a text-to-text translation system without using any paired data [86, 87, 88, 89]. However, those works focus on mapping within a single modality framework (i.e., speech-to-speech or text-to-text). On the other hand, mapping between different modalities is more challenging due to the differences in the data characteristics.

In a speech-to-text mapping task, speech features are continuous vector sequences while the corresponding text is formed in discrete sequences. Unfortunately, scant research has considered multi-modality mapping tasks with non-parallel data. Within the limited research on speech-to-text mapping tasks with non-parallel data, Sarl et al. (2020) recently proposed a spoken language understanding system trained on non-parallel speech and text data [90]. However, the model is more focused on dialog-act recognition rather than generating a descriptive sentence.

In this chapter, we focus on generating a descriptive sentence of the message being spoken. Specifically, the system attempts to learn how to generate semantically related text messages from speech utterances. We introduce the possibility of conducting weakly-supervised learning based on non-parallel data using a partially-aligned Transformer. We also introduce discrete speech representations using a vector-quantized variational autoencoder (VQ-VAE) to reduce the complexity of speech-text mapping, which also solves the low-resource problem and opens up possibilities for our proposed method to be used in an untranscribed unknown language.

5.3. Proposed Weakly-supervised Speech-to-text Mapping

Our proposed framework transforms a speech X into a sentence Y, by leveraging the non-parallel speech and text data (Figure 5.1). We use part of the MMC general framework as described in Section 3.2.5. First, to simplify the speech variability and its length discrepancies in text, we train a Transformer-based VQ-VAE to learn a discrete speech representation in a self-supervised manner. Then, we perform unsupervised alignments between the resulting discrete speech representation and the discrete target text sequences. Since the speech source and target text are generally based on the same images, we assume that some parts of speech and text content are semantically associated or aligned, which are then used by a partially-aligned Transformer model for speech-text mapping. Finally, we use the cycle mechanism as an augmentation to further improve the partially-aligned Transformer model.

5.3.1 Model Components

- **Transformer-based Vector-quantized Variational Autoencoder**
  We use the VQ-VAE model with speaker embedding described in Section 2.1.3 to learn speech discrete representation in a self-supervised manner.

- **Partially-aligned Code2Text Transformer Model**
  A partially-aligned Code2Text model uses the alignment of discrete speech representation $C = \{c_0, c_1, \ldots, c_n\}$ with the discrete target text sequences $Y = \{y_0, y_1, \ldots, y_m\}$. Inspired by the partially-aligned training strategy [91] for sequence-to-sequence neural machine translation (NMT), we modified a vanilla Transformer-based NMT model [62] into a partially-aligned Code2Text Transformer model by leveraging the alignment information between the input and output (see Figure 5.2). Let us assume that P_c and P_y form the list of aligned words from the C and Y sequences. First, we penalized the source-to-target attention score in the decoder, so if $y_j \notin P_y$, the attention context vector for that word is zero ($C_t = 0$). Then, we also add an additional attention loss to emphasize the alignment between the partially-aligned part in a supervised manner. We create a hard-attention matrix H, where:

$$H_{i,j} = \begin{cases} 1 & \text{if } c_i \in P_c \text{ and } y_j \in P_y \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

  so that the original attention matrix A can be supervised with attention loss L_{att} as follows:

$$L_{att} = \sum_{i=0}^{n} \sum_{j=0}^{m} \|A_{i,j}, H_{i,j}\|_2^2. \tag{5.2}$$

  Finally, we weighted the softmax cross-entropy loss L_{ce} with L_{att} as follows:

$$L = L_{ce} + \alpha L_{att}. \tag{5.3}$$

- **Cycle Mechanism**
  We implemented a Code2Text and Text2Code chain to further improve the

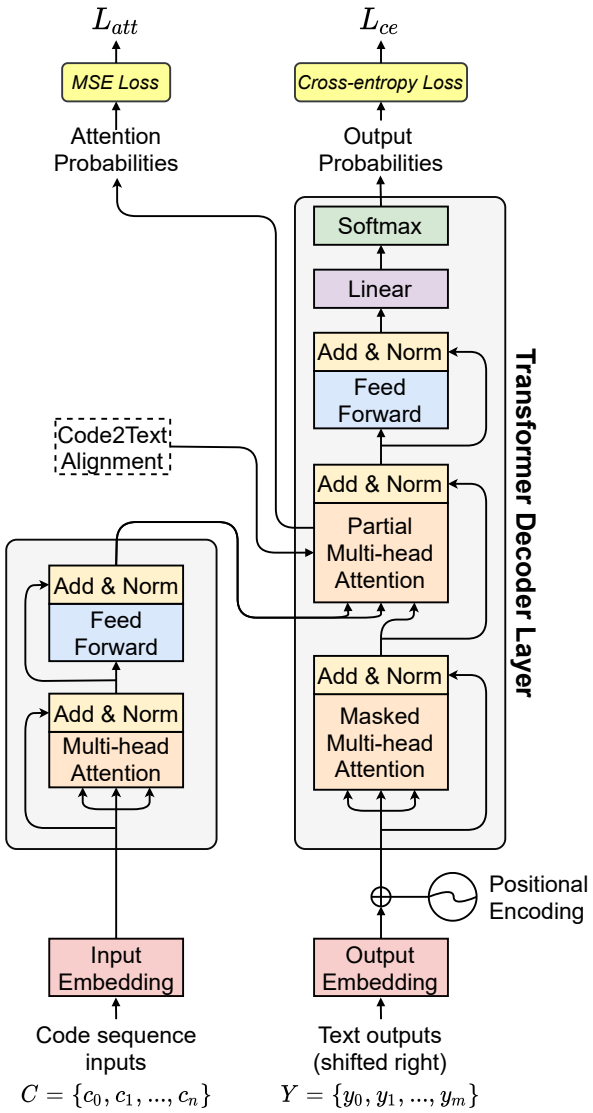

Figure 5.2: A Transformer-based Code2Text for partially-aligned input-output

performance of our proposed method (Figure 5.3). We used the chain training mechanism for unpaired data described in Section 3.2.3, but without using any paired data. Given a text-only dataset D_y, a text y is translated using the Text2Code model, generating a $\hat{c}$ code hypothesis. This code hypothesis is then translated back into $\hat{y}$ by the Code2Text model. Then, we can backpropagate the Code2Text model using the reconstruction loss between y and $\hat{y}$.

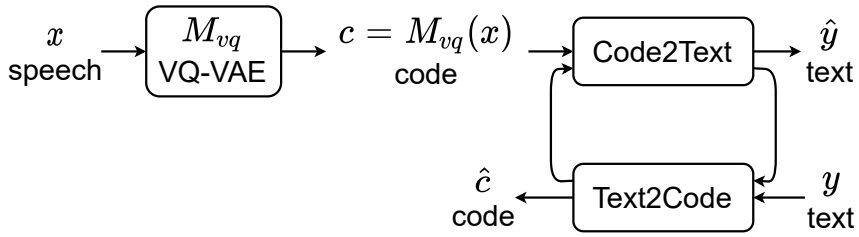

Figure 5.3: Unsupervised augmentation with chain mechanism

5.4. Experiment Settings

5.4.1 Visually-connected non-parallel speech-text data

Common parallel dataset:

$= \{(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\}$

Visually-connected non-parallel dataset:

$= \{(x_0, y_2), (x_1, y_3)\}$

Figure 5.4: Visually connected non-parallel speech (x) - text (y) data

We used Flickr8k [71], which contains 8k images of everyday activities and events. For the synthetic speech caption, we generated single-speaker speech using GoogleTTS from the text caption. Then, for the natural speech caption, we used Flickr Audio [92], which was recorded using the crowdsourcing method with 183 unique speakers. We use the development and test sets which consist of 1k images each.

We formulated the training set differently from the original Flickr8k dataset by splitting the data as a visually grounded paraphrase (VGP) [93] to ensure a "semantic equivalence". In this task, we need to show how our proposed method can learn from non-parallel speech-text data but with semantically similar meaning. While each image in this dataset has five speech and text captions, we choose two captions as speech only data and another two captions as text-only data (Figure 5.4). Therefore, both the speech and text have the same image, which guarantees semantic equivalence between the pseudopair. This partition

yields 12k speech utterances, and 12k of text caption, with both sets are disjointed $((x_i, y_j), i \neq j)$.

To show that our proposed approach can also be applied for a cross-lingual application, we also ran a cross-lingual experiment using English speech from SpeechCOCO multispeaker dataset [6] to the Japanese text from the STAIR Caption dataset [7], where this non-parallel speech-text mentions the same image from the MSCOCO dataset [94]. We take the matching amount of data and process it similarly to how the Flickr8k dataset is handled. We assume that the English data is the speech from an unknown and untranscribed language, where a visually connected non-parallel Japanese text exists.

5.4.2 Model Parameters

We extracted the Mel-spectrogram (80 dimensions, 25-ms window size, 10-ms time steps) using the Librosa package [95]. This speech feature is used as the input and output of the VQ-VAE model that has a 256 codebook size and 32 code dimensions. For adapting with the natural speech dataset, we froze the codebook part of the VQ-VAE model so that each code still represented the same speech segment. We trained the VQ-VAE model with the bucket size of 20000 frames, with 150 epochs in average. We choose the model in the epoch with the lowest development loss to generate the discrete speech representation.

We used a Transformer-based text encoder and decoder with a depth of 6 and a size of 512 hidden units. For the output layer, we used label smoothing with a factor of 0.005 and beam decoding with a size of 3. The vocabulary consists of words in the text-only training data that appear at a frequency of more than one time. We used Fast Align [96] as the unsupervised aligner. The Code2Text and Text2Code model consist of 6 transformer layer each for encoder and decoder, using multihead attention with the size of 512 and a feed forward layer with the size of 2048. We used dropout mechanism with the probability of 0.2 for embedding layer, and 0.3 for the transformer layer. Both the Code2Text and Text2Code model are trained with the batchsize of 50. Inside the chain mechanism, we only updated the last element of the chain due to memory limitation. We trained all models with the Adam optimizer [73] using a learning rate of 1e-4. We halved the learning rate when the development loss does not decrease after a warmup of

five.

5.4.3 Evaluation Method

We evaluate our proposed model performance by running an inference step using the dev and test set of the dataset the model is trained with. We use common metrics in the image captioning task: bilingual evaluation understudy (BLEU) with 4-gram [76] and CIDEr [97]. A BLEU score measures the n-gram similarity between the hypothesis and references, while CIDEr measures consensus by evaluating beyond the n-gram exact similarity. We used both metrics in a multi-reference condition. In addition, to evaluate the semantic aspect, we developed a cosine-similarity based metric (Sim%) for multi-reference evaluation by calculating the highest cosine similarity between the hypothesis sentence embedding and the reference sentence embeddings. We generated the sentence embedding using the Sentence Transformers toolkit [98] with the pretrained models of RoBERTa [99] for English and Universal Sentence Encoder [100] for Japanese.

We calculated the corpus vocabulary statistics such as the number of unique words and the vocabulary utilization ratio to measure how rich the hypotheses are. We reported the Pearson's correlation coefficient (r) score between the word frequencies of the hypothesis and the training set to show how good a model could learn to mimic the training set's word distribution.

5.5. Experiment Result and Analysis

5.5.1 Result on Single-speaker Synthesized Speech Non-parallel Dataset

In Table 5.1, we provide the baseline score of a random selection to show that our trained model produces a coherent hypothesis. We also reported the score of the ASR model trained directly on the non-parallel speech-text data. Then, we trained the VQ-VAE model using the speech data, and generated the code sequence as a discrete speech representation. The code sequence can then be used to train a Code2Text model against the partially-aligned text caption. Our proposed Code2Text model delivers a better score than the ASR baseline, which shows that

Table 5.1: Experiment result in the Flickr8k synthesized speech non-parallel dataset

Model	Dev			Test		
	Sim%	BLEU	CIDEr	Sim%	BLEU	CIDEr
(Baseline)						
Random selection	16.73	2.28	3.42	16.25	2.22	3.58
ASR [21]	16.86	5.64	7.63	16.94	4.69	7.08
(Proposed)						
Code2Text	35.58	15.30	31.48	35.79	15.04	31.66
+Partial Code2Text	40.58	16.95	36.11	40.94	16.80	36.86
+Cycle Augmentation	40.03	16.74	36.44	**40.47**	**17.25**	**37.52**

our discretization method using VQ-VAE provides more efficient learning due to reduced variability compared with mel-spectrogram.

Then, because the input and output are discrete, we can approximate the alignment between the generated code sequence and the partially-aligned text using an unsupervised aligner. We next use the alignment information to influence the source-to-target multi-head attention by producing an additional L_{att}. We found that by multiplying L_{att} with $\alpha = 0.9$, we could obtain about 5.15% cosine similarity and 5.2 CIDEr points improvement on the test set, compared with a no-alignment model (Code2Text). We also trained the partially-aligned Text2Code with the same steps. After that, we use it in a cycle mechanism to achieve cross-modal augmentation which yielded a 0.66 CIDEr improvement. We also trained an ASR model with parallel data for a topline comparison, which yield 89.81% cosine similarity, 81.43 BLEU, and 206.59 CIDEr scores on the test set.

5.5.2 Adaptation Result on Multispeaker Natural Speech Non-parallel Dataset

Furthermore, we adapted our trained model to also support a multispeaker natural speech dataset using the Flickr8k multispeaker natural speech dataset, which we also use for testing. As shown in Table 5.2, our adaptation improves CIDEr by 20 points compared to the baseline ASR and 17 points compared to simply using the best model in Table 5.1 (no adaptation). We also trained a topline model with the parallel dataset, which yields 82.75% cosine similarity, 70.24 BLEU, and 176.42 CIDEr scores. Next, we took the best score of the test set from Table 5.2

Table 5.2: Adapting best Speech2Text model trained on Table 6.2 to the Flickr8k multispeaker natural speech non-parallel dataset

Model	**Dev**			**Test**		
	Sim%	**BLEU**	**CIDEr**	**Sim%**	**BLEU**	**CIDEr**
(Baseline)						
ASR [21]	16.30	3.23	9.30	15.18	3.09	9.07
(Proposed)						
Cyclic Partial Code2Text						
no adaptation	21.37	7.84	11.64	21.31	7.83	11.69
with adaptation	35.70	14.64	29.80	**35.35**	**14.57**	**29.01**

Table 5.3: Our proposed Speech2Text vocabulary utilization statistics for the Flickr8k multispeaker natural speech dataset (Table 2) in comparison to the baseline.

Metric	**Baseline**	**Proposed**
Number of unique words	20	300
Vocab utilization ratio	0.69%	10.42%
Pearson correlation (r)	0.343	0.958

and compared the corpus statistics in Table 5.3. We found that the baseline system did not converge, as shown by the very low number of unique words with only 0.69% of the vocabulary being used. In comparison, our proposed model yielded 10.42% vocabulary utilization ratio. Moreover, our proposed method shows better modelling of the vocabulary with a Pearson correlation (r) of 0.958, which is close to the topline of 0.999. This shows that our proposed partially-aligned Code2Text can model the training set word distribution as successfully as the topline, even without using any parallel data. In addition, even with limited vocabulary, our proposed method can still effectively convey the semantics of a partially-aligned speech.

Table 5.4: Example results from the test set

Model	**Sentence**
Baseline	two dogs are running through the grass .
Proposed	a woman and a little girl are smiling .
Reference	a laughing woman holding a little girl .
Baseline	a man and woman pose for a picture .
Proposed	a man in a red shirt is rock climbing .
Reference	a man poses as he jumps from rock to rock in a forest .

Table 5.4 shows a comparison of results between our proposed model, baseline ASR, and the input speech transcription (reference). The first example shows the

baseline ASR model hypothesis which is totally unrelated to the reference. Our proposed method generated a hypothesis that semantically, closely resembles the reference, even while replacing the word "laughing" with "smiling". Then, in the second example, our proposed method successfully described the rock-climbing activity mentioned in the speech (reference). Although it is not an exact one-to-one transcription, the speech content itself can be successfully described in each of our proposed method's generated hypotheses. We are confident that this result will be very useful under the condition where no parallel speech-text data are available, in addition to handling an untranscribed unknown speech language.

5.5.3 Result on Cross-Lingual Scenario

Table 5.5: Experiment result under cross-lingual EN-JA condition of transforming multispeaker English speech [6] to non-parallel Japanese text [7]

Model	**Dev**			**Test**		
	Sim%	**BLEU**	**CIDEr**	**Sim%**	**BLEU**	**CIDEr**
ASR [21]	24.85	2.39	1.63	25.13	2.50	1.54
(Proposed)						
Code2Text	30.15	13.17	12.96	30.28	13.49	13.22
+Partial Code2Text	30.08	13.33	13.94	30.06	13.41	13.57
+Cycle Augmentation	30.51	13.36	14.21	30.33	13.40	13.75

Finally, we demonstrate how our proposed method can be used under a cross-lingual condition. As shown in Table 5.5, we found that the partial Code2Text and the cycle augmentation showed a little improvement in terms of CIDEr score. We hypothesize that this is due to the difficulty of aligning between different language structures (i.e., SVO for English, but SOV for Japanese). Nevertheless, while the baseline ASR did not show convergence, our proposed model could still achieve BLEU score of about 13 points even with a small amount of non-parallel data. This shows the effectiveness of our proposed discretization using the transformer-based VQ-VAE.

5.6. Summary

In this study, we use our proposed MMC framework for a weakly-supervised mapping task to transform unknown untranscribed speech utterances into a semanti-

cally equivalent text, even without a parallel speech-text dataset. Our proposed system uses a pipeline of VQ-VAE to generate a discrete speech representation, and a partially-aligned Code2Text Transformer model to learn the mapping between the code and the text. We also employed a cyclic augmentation strategy to further improve the performance of the Code2Text model. Our experiments with a multispeaker natural speech dataset showed improvement in every aspect that we examined. Our analysis of the text hypothesis shows that our proposed method can produce a more semantically relevant text. For future work, we will explore methods to increase the vocabulary utilization ratio, including an adversarial training method.

Chapter 6

MMC Framework for End-to-end Image-to-speech Generation for Untranscribed Unknown Language

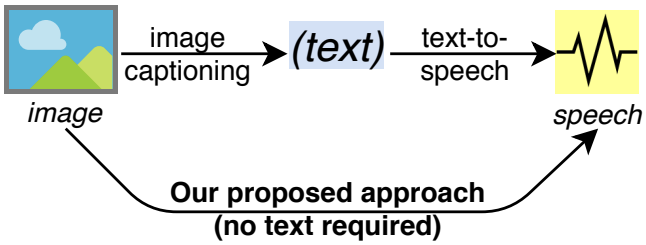

Figure 6.1: Image2Speech: direct image-to-speech captioning without using text as a bridge.

This chapter describes our approach to create synthesize a speech caption from an image, using part of the MMC framework defined in Section 3.2.6. As commonly known, speech has a continuous representation with a high variability due to the difference in pronunciation due to various factors such as speaker, context, and others. To reduce this variability so that a direct image-to-speech is possible, we introduce a discretization step using part of our MMC framework that supports **self-supervised learning**, which we call as MMC-SelfSup. This

chapter describes our attempt to use the MMC-SelfSup as a part to develop an end-to-end Image2Speech system that does not need any textual information in its training.

6.1. Introduction

Popular natural language technologies such as automatic speech recognition (ASR), machine translation (MT), and image captioning (IC) are mostly built on the assumption that every language has an orthographic representation. However, some languages do not have reliable orthographic features such as those that can provide a textual transcription [101], which renders these technologies impractical. To overcome this issue, an approach named "Zero Resource Speech Technology" aimed to construct a speech system without any textual representation [54, 102, 103]. This kind of setting is inspired by the fact that young children can communicate effectively in their native language even before they learn to read or write [54].

Several applications of this zero resource setting have been explored in building a speech representation for tasks that mimic human language development such as acoustic unit discovery [104, 105, 106, 107, 3], subword modelling [108, 109, 110], and spoken term discovery [111, 112, 113]. These tasks focused on discovering a speech representation with minimal supervision, which generalizes across languages, especially for an unseen language. On the speech production side, the Zero Resource Speech Challenge 2019 aimed to build a TTS without T (text-to-speech without text) to bolster the zero resource settings in terms of speech generation [102, 3]. Moreover, there was also an attempt to achieve speech-to-speech translation between untranscribed unknown languages [114].

Table 6.1: Our contribution in comparison with Hsu et al. (2020) [19]

Factor	**Hsu et al. (2020)**	**Ours**
Speech representation model	Visual grounding-based	Speech reconstruction-based with VQ-VAE
➛ *Training data*	Required a large amount of paired image-speech data	Possible with speech-only data
➛ *Duration information*	Information lost due to run-length-encoding (RLE)	Information intact because of frame-based encoding
Speech generation model	Required to train separate TTS with additional speech data	Possible without additional TTS and additional data. Generate speech from the speech representation model.
Optimization	Separately	End-to-end finetuning

Given these attempts to develop language technologies for these languages, there is one crucial aspect of natural human communication that has not yet been explored. Orally describing what we are seeing is a simple task that we are able to do from our early days. This activity includes an observation of what objects are seen and what are the relationships among those objects. Then, we speak to describe the observed object in the form of a descriptive sentence. Currently, the construction of such a system is done by cascading image captioning and TTS models, which come from two separate fields: image and speech processing. Unfortunately, this implementation still relies on the textual modality as a bridge, which makes it impossible for use in untranscribed unknown languages. In this study, we combine these two fields by proposing an end-to-end direct Image2Speech generation method without text (see Figure 6.1), which we call MMC-SelfSup.

Consequently, the simplest way to accomplish this Image2Speech task is to have a large amount of parallel image and speech data to train a generation model. However, such parallel data are not common as data pairs and thus are often unavailable. This has not yet been addressed in the previously published Image2Speech works [115, 116, 5].

From these observations, we find that there are two main problems: (1) some languages do not have a written form and (2) a large amount of parallel image-speech data are often unavailable. To resolve these problems at once, in this study, we use a simpler approach to learning a discrete speech representation from speech-only data, instead of using the grounding-based approach [5] that always needs paired image-speech data. We use a self-supervised transformer-based vector-quantized variational autoencoder (VQ-VAE), which has been proven to deliver a promising discretization score for untranscribed unknown languages in the recent Zero Resource Speech Challenge [107, 3]. As a result, we greatly reduce the amount of needed parallel image-speech data, as shown by our experimental results that surpass those of the most recent frameworks, even while using a smaller amount of parallel data.

6.2. Related Work

6.2.1 Image2Speech with Text

Image2Speech is a relatively new task that bridges image and speech processing. Previously, both modalities have been processed separately using textual modalities as a bridge between them. This same approach was done by Ma et al. (2019) by learning multimodal representations [117]. With their proposed multimodal information bottleneck framework, they were able to train a model with disjointed image-text and text-speech datasets. Then, during the generation process, they did what they called a "skip-modal generation" so that the shared modality (text) is skipped during the generation process. Although this work successfully captions an image into speech, its training process still uses text as a bridge, which is different from what humans do during their early learning process.

Hasegawa-Johnson et al. (2017) proposed a more direct approach for Image2Speech tasks by converting an image feature into speech unit sequences such as L1-Phones and L2-Phones. The speech unit sequences are then used by TTS as inputs to generate speech utterances. In addition, they also experimented with pseudo-phones, which are generated by an unsupervised statistical-based acoustic unit discovery system [104]. However, their approaches using L1- and L2-Phones still need textual information to generate the phones from the speech. Then, their pseudo-phones approach performed poorly with 1.4 BLEU on a synthetic speech dataset and it was not tested on a natural speech dataset.

6.2.2 Image2Speech without Text

In the previous section, those previous studies used text as a symbolic representation of the objects found in the image. Similarly, to implement a text-free approach, Hsu et al. (2020) proposed a visual-grounding approach [5] that generates a set of discrete units associated with speech segments and visual objects. Then, they trained a TTS model using a separate speech-generation dataset with the input of those speech units. However, although this approach can successfully replace text as an intermediate representation, the training process for discrete

speech units still needs a full amount of parallel image-speech data, which are not commonly found data pairs in real-life conditions. In addition, they also applied run-length encoding (RLE) so that each speech unit's discrete representation could be easily assigned to an image object. Unfortunately, RLE is a lossy approach that removes the duration information from the speech units, and thus a separate TTS model is needed to recover the duration information during the unit-to-speech generation process.

In our proposed approach, we train the discrete speech unit representation with a speech-only dataset in a self-supervised manner, as opposed to the approach of Hsu et al. [5] that needs paired image-speech data (See Table 6.1). This consequently reduces the amount of parallel image-speech data needed in our proposed framework. In addition, our reconstruction-based autoencoder model can be re-used during the inversion of a speech unit into speech. In this way, we are also able to remove the use of the previously required lossy RLE (e.g. [5]), which is impractical because it removes duration information from the speech. Therefore, we do not need to train a separate TTS model as did other approaches (e.g. [115, 116, 5]) because the speech-generation process from the discrete unit is simply an inversion process in the autoencoder part of our proposed framework. Moreover, our proposed approach also does not need any additional speech synthesis dataset for training a TTS model, as needed by another work [5]. As a result, we propose an architecture for the Image2Speech task without any text, where the data requirement for training also corresponds to real-life conditions (i.e. lots of speech-only data, few parallel image-speech data).

6.3. Proposed Self-supervised Discrete Speech Representation for End-to-end Image2Speech Generation

6.3.1 Model Components

- **Vector-quantized Variational Autoencoder (VQ-VAE)**
  We use the VQ-VAE model with speaker embedding described in Sec-

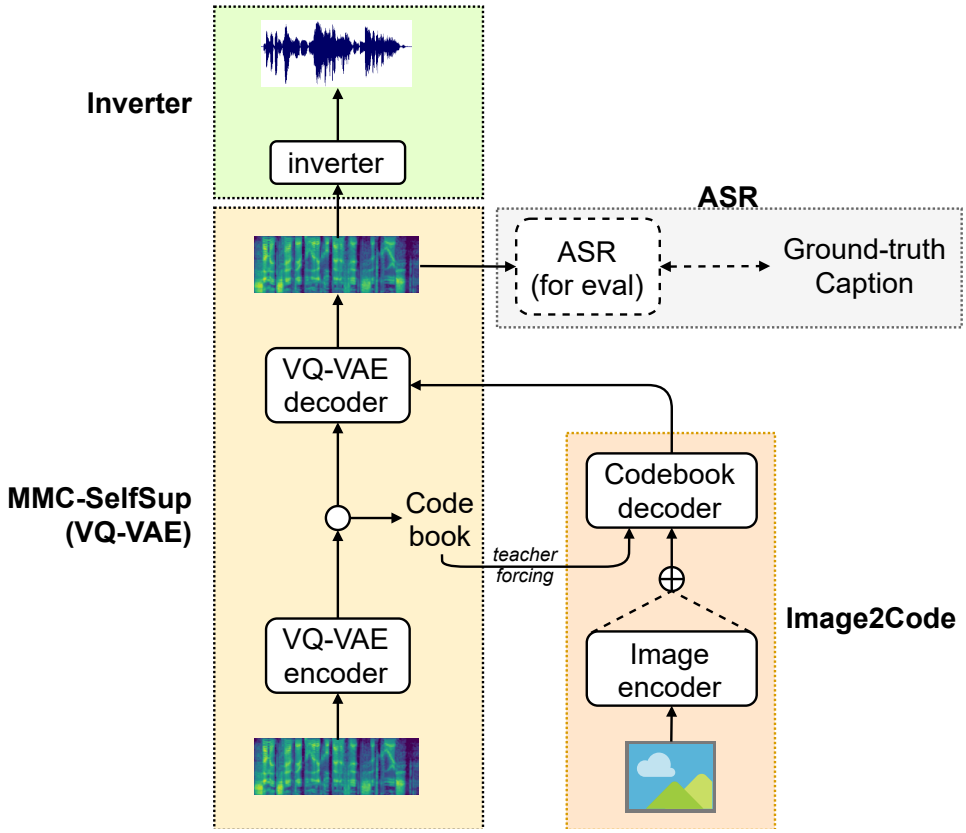

Figure 6.2: Overview of our proposed Image2Speech for direct image-to-speech captioning without text.

tion 2.1.3 to learn speech discrete representation in a self-supervised manner. We call this MMC-WeakSup, based on the MMC framework definition in Section 3.2.6. We can list the variables for this task as follows:

- VQ-VAE represents a chain path $\mathcal{C}_{XX^DX}$, where
- modality X is speech, and
- learned representation X^D is the codebook of the VQ-VAE.

- **Transformer-based Image2Code Generation**
  We use a transformer-based image captioning model mentioned in Section 2.2.1. We train the text decoder part using teacher forcing on the sequence of codebooks C. This determines the size of the output-embedding

and softmax layer on the end, which is constrained to the number of codebooks predefined during the VQ-VAE training instead of the actual vocabulary number.

- **Mel-spectrogram inverter**
  We use the inverter part of the TTS model described in Section 2.1.2 to invert the generated mel-spectrogram sequence into a speech waveform in two steps. The first step is the inversion of this Mel-spectrogram into a linear-spectrogram sequence. For this task, we used the inverter part of the Tacotron TTS [2], which consists of a 1-D Convolution Bank + Highway + bidirectional GRU (CBHG) block, followed by a fully connected layer. We used L2 loss to compare the predicted linear-spectrogram with the original linear-spectrogram extracted from the speech. The second step is to use the Griffin-Lim algorithm [56] to iteratively estimate the phase spectrogram so that the waveform can be reconstructed with the inverse short-time Fourier transform (STFT).

6.3.2 Multispeaker Natural Speech Adaptation

In this study, we want our proposed model to also be able to generate multi-speaker speech. However, using a multi-speaker natural speech dataset poses a challenge to the VQ-VAE discretization due to its vast pronunciation variation. In this study, we first trained the VQ-VAE model using the Flickr8k synthesized dataset and measured its performance in the Image2Speech pipeline. Then, we adapted the VQ-VAE model with the multi-speaker natural speech Flickr8k dataset in the form of fine-tuning. During the fine-tuning, we froze the codebook so that each codebook still represented the same speech unit as before. In this way, our VQ-VAE model can reconstruct natural speech with a multi-speaker condition while maintaining a meaningful codebook representation.

6.3.3 End-to-end Model Integration

The codebook hypotheses generated by the VQ-VAE model might contain some errors. Since the Image2Code model is trained using teacher forcing against this codebook hypothesis, the error from the codebook selection is also propagated to

the final speech-generation process. To solve this problem, we entirely skip this codebook step by combining the Image2Code model and the VQ-VAE model in an end-to-end fashion. We integrate the output layer of the Image2Code model with the decoder part of the VQ-VAE model.

We define the end-to-end loss L_{e2e} as a weighted sum of the Image2Code cross-entropy loss L_{CE} with the VQ-VAE reconstruction loss L_{recon} as follows:

$$L_{e2e} = \alpha L_{CE} + \beta L_{recon}. \tag{6.1}$$

To connect the output layer of the Image2Code model with the VQ-VAE decoder, we multiply the posterior probability vector p_i for each codebook with each codebook vector itself. In this way, the codebook selection ambiguity can be reflected in the VQ-VAE decoder.

$$\hat{z}_i = \sum_{i=1}^{L} p_i c_i. \tag{6.2}$$

.

6.4. Experiment Settings

6.4.1 Dataset

We used Flickr8k [71], which contains 8k images of everyday activities, scenes, and events, with five captions each. For the synthetic speech caption, we generated single-speaker speech using GoogleTTS from the text caption. Then, for the natural speech caption, we used Flickr Audio [92], which was recorded using the crowdsourcing method. There are 183 unique speakers in this dataset. The training, development, and test sets consist of 6k, 1k, and 1k images, respectively.

6.4.2 Experiment Settings

We extracted the speech feature of Mel-spectrogram (80 dimensions, 25-ms window size, 10-ms time steps) using the Librosa package [95]. For the inverter, we also generated a linear magnitude spectrogram with 1025 dimensions for the

Mel-spectrogram inverter. For images, we used a 224×224 image size with augmentation (dynamic resize, random crop, random flip). All of our models are trained with the Adam optimizer [73] with 1e-4 learning rate for VQ-VAE and 2.5e-4 for image-captioning and inverter.

We build our VQ-VAE with a transformer layer having a depth of three for each encoder and decoder part. We use multi-head attention with the size of 256 and feed forward layer with the size of 1024. In addition, we also used an embedding layer to represent a speaker id auxiliary input into the speaker embedding with the size of 8. The speaker embedding will be concatenated to each element of the decoder input sequence as the additional reconstruction condition. When the training failed to reduce the development loss after a warmup of five, we halved the learning rate. The training stops when the development loss does not decrease anymore, with the average of 150 epochs. The training was done with a bucket size of 20000 frames.

Our Image2Code uses a pretrained ResNet-50 [4] as an image encoder, which high-level feature of 14×14×512 dimension is then attended by a transformer-based text decoder with a depth of 6 and a size of 512 hidden units. For the output layer, we used label smoothing with a factor of 0.005 and greedy-based decoding unless specified. We use the same halving mechanism for the learning rate as the VQ-VAE model. We use 32 captions per batch during training, which are shuffled beforehand.

Finally, to listen to the generated speech caption, we inverted the Image2Code code output into Mel-spectrogram. Then it is converted into a linear-spectrogram by the inverter model with a projection size of 256, which then finally transformed into a waveform by the inversion process described in Section 6.3.1. Since the latter part of the inversion process (Mel-to-speech) is just to show the possibility of inverting our generated Mel-spectrogram into a speech waveform, we also encourage the use of other waveform synthesizer models.

6.4.3 Baseline and Topline Model

As a simple baseline, we combined two popular neural network models: ResNet-50 [4] and Tacotron [2]. We used ResNet, similarly to our Image2Code model, as an image encoder that produces two-dimensional image features. Then, we take

the decoder part of Tacotron so that it attends to these two-dimensional image features. The Tacotron decoder is then trained using teacher forcing against the Mel-spectrogram speech feature from the image's caption. In addition, we compare the proposed approach's result with that of Hsu et al. [5], which used a visual grounding-based model.

We also trained another baseline that leverages international phonetic alphabet (IPA) as an intermediary to enable transfer learning from another language. IPA is an international phonetic notation that represents speech sounds in written form. To train this baseline system, we assume that there exists a dataset in another language that has textual information. We generated the German speech from the German image captions in Multi30k dataset [118] using GoogleTTS. The transfer learning flow from German to English is as follows:

- First, we generated IPA transcriptions of German based on the textual caption and pronunciation dictionary using Epitran[1] transliteration tool.
- Second, we trained automatic speech recognition (ASR) using the pair of German speech and its IPA transcription (Speech2IPA).
- Next, we transcribed the English speech in the Flickr8k dataset using this Speech2IPA model, so an English image captioning model (Image2IPA) can be trained using the generated IPA transcription.
- Then, the parallel of speech and IPA transcription in English is used to train a TTS model (IPA2Speech).

During inference, given the image, the English caption (in terms of IPA) will be generated by the Image2IPA model. Then the IPA2Speech model will produce the speech based on the IPA caption in English. This baseline evaluation can be used to measure the significance of our proposed discrete speech representation compared with borrowing such features from another language.

For the topline, we follow the commonly used pipeline to make a speech caption out of an image by generating the text caption beforehand. First, we trained the Image2Code model to generate phoneme captions. Then, we also train a

[1]https://github.com/dmort27/epitran

Tacotron [2] model to generate speech from phoneme sequences as a topline comparison. During inference, the Image2Code produces phoneme caption hypotheses, which are then used by the Tacotron to generate speech captions. All of the models for the topline and baseline were trained with the same Flickr8k dataset as our proposed model.

Then, we also consider the use of the previously published work of Hout et al. [116] that uses L1-Phones as a topline, since the training process still needs textual information. However, the earlier work of Hasegawa-Johnson et al. [115] cannot be compared because they used a different evaluation method as previously mentioned [116]. Their proposed system was also similar in principle to that of Hout et al., so we assume that their result is also representative.

6.4.4 Evaluation

We used phoneme-level and word-level granularity to evaluate the output of our system, following the approaches of Hout et al. (2020) [116] and Hsu et al. (2020) [5], respectively. We used the common metrics for image captioning task: bilingual evaluation understudy with 4-grams (BLEU4) [76] and CIDEr [97]. Both metrics are used in phoneme level for our proposed model parameter tuning and for comparison with Hout et al.'s result. BLEU4 is used as the main metrics because it was reported with the strongest correlation with human evaluators for this Image2Speech task [116]. Then, for comparison with Hsu et al.'s result, we also used METEOR [119] and ROUGE-L [120]. As commonly done in the image-captioning field, we evaluated our model hypothesis with multiple references (each image has five captions).

To generate the phoneme transcription, we trained a sequence-to-sequence automatic speech recognition (ASR) model based on the "Listen, Attend, and Spell" (LAS) framework [21] with the reconstructed Mel-spectrogram of the VQ-VAE as input and phoneme or word (textual) caption as the transcription output. Using the textual transcription of the generated Mel-spectrogram transcribed by this model, we compared it with the ground truth text caption from the dataset. This evaluation approach is similar to that of a previous work [5].

6.5. Experiment Result and Analysis

There are several parameters such as the number of clusters, cluster size, and stride in the proposed model that need to be evaluated. In this section, we report the results of our parameter tuning and how each parameter affects the final performance.

6.5.1 Result on Single-speaker Synthesized Speech Dataset

First, we compared the number of clusters (codebook size) for the VQ-VAE part of our proposed model. There is a trade-off between the quality of the VQ-VAE reconstructed speech and the number of vocabulary items used by the Image2Code module. As shown in Table 6.2a, we found that the ideal number of clusters is 256, which yields a 36.29 BLEU4 score. Therefore, we chose this setting for the next experiment to choose the cluster size. We found that the increase in cluster number is not always positively correlated with the end performance because after convergence the VQ-VAE model did not use all of the possible codebooks to represent the given speech-caption utterance.

Then, if we regard each code as a speech unit, the cluster size parameter represents the dimensions needed to represent each of those units. A bigger cluster size gives a richer representation, which we expected to give a better codebook sequence representation and better reconstructed speech quality. However, the results in Table 6.2b show that a cluster size of 32 is good enough to represent the speech unit generated by the VQ-VAE model. A further increase in the cluster size did not further improve performance and made the system prone to overfitting. We suspect that this is due to the size of the dataset used, which is not large enough.

Another factor that needs to be evaluated is how wide the stride size must be. Stride size represents how many frames are represented by each codebook. A smaller stride size produces a better fine-grained codebook representation but makes the codebook sequence longer. Such a longer codebook sequence poses more difficulties for the Image2Code module training. On the contrary, a bigger stride size provides a more robust codebook representation because it spans a wider receptive field. We found that the stride size of 8 produces 0.88 more

Table 6.2: Experiments on Flickr8k single-speaker synthesized speech dataset. (phoneme-level evaluation)

(a) Comparing number of cluster

#code	**size**	**stride**	**BLEU4 ↑**	**CIDEr ↑**
64	32	4	33.98	40.28
128	32	4	35.35	45.53
256	32	4	**36.29**	45.88
512	32	4	35.46	46.95
1024	32	4	35.16	43.84
2048	32	4	34.17	44.71

(b) Comparing cluster size

#code	**size**	**stride**	**BLEU4 ↑**	**CIDEr ↑**
256	16	4	35.96	45.38
256	32	4	**36.29**	45.88
256	64	4	35.23	46.63
256	128	4	34.17	40.99
256	256	4	35.26	43.99
256	512	4	34.17	42.90

(c) Comparing various strides

#code	**size**	**stride**	**BLEU4 ↑**	**CIDEr ↑**
256	32	2	32.88	38.56
256	32	4	36.29	45.88
256	32	8	**37.17**	49.23
256	32	12	36.29	42.33

(d) End-to-end integration between Image2Code and VQ-VAE decoder. Pre-trained model taken from best model in Table 6.2d. α: Image2Code loss weight, β: reconstruction loss weight

α	β	**BLEU4 ↑**	**CIDEr ↑**	**Notes**
0	1.0	1.64	0.02	reconstruction loss only
any	0	37.17	49.23	before end-to-end integration
1.0	0.25	37.12	48.31	
1.0	0.50	36.91	46.62	weighted reconstruction loss
1.0	0.75	**37.78**	49.07	
1.0	1.0	37.43	47.54	equal for both

BLEU4 points than the stride size of 4 (see Table 6.2c).

Then, we used the best settings for the end-to-end fine-tuning between Image2Code and the VQ-VAE decoder. We connect the output layer of Image2Code

to the embedding input layer of VQ-VAE as described in Section 6.3.3. Table 6.2d shows the effect of end-to-end fine-tuning applied on the best model of Table 6.2c, with various combinations of α and β parameters. We regard the result when β=0 as the same as that with no integration because the Image2Speech system relies only on cross-entropy loss from the Image2Code model. The results show that by using 0.75 multipliers for the β parameter, we can get 0.61 BLEU4 improvements. A comparison of the end-to-end integration results with no integration is shown in Figure 6.3. In addition, we found that simply relying on reconstruction loss (α=0,β=1) reduces performance greatly.

6.5.2 System Adaptation to Multi-speaker Natural Speech

Table 6.3: Adaptation results on Flickr8k **multi-speaker natural speech** dataset (phoneme-level evaluation)

#code	**size**	**stride**	**BLEU4** ↑	**CIDEr** ↑	△**BLEU4**↑
256	32	2	32.76	38.48	-0.12
256	32	4	36.40	45.87	+0.11
256	32	8	37.26	49.29	+0.09
256	32	12	**37.73**	44.09	+1.44

The model was adapted so it could be trained with a multi-speaker natural speech in the form of fine-tuning, as described in Section 6.3.2. By incorporating the speaker information as an additional condition in the VQ-VAE decoder, we can adapt the entire pipeline to also generate multi-speaker natural speech. Table 6.3 shows the result of the fine-tuning adaptation. The best performance was reached by using a stride size of 12, different from the one in the synthesized speech. We also calculated the △BLEU4 score in comparison with the synthesized speech results in Table 6.2c. Interestingly, we found that a stride size of 12 is better at handling multi-speaker natural speech, as compared with the optimal stride size in the synthesized speech model. This validates our hypothesis that a larger stride size provides a more robust representation against diverse pronunciation variation, which is typical with multi-speaker natural speech data.

We added the example results of this model in Figure 6.4. The generated speech output hypothesis of our proposed Image2Speech model is transcribed by

an ASR model for evaluation. Example results (1) and (2) are good results, where the object in the images is correctly captioned alongside its action. Example result (3) is logically correct but sounds unnatural because "a group of people" is not a commonly used phrase to describe two people. Example result (4) is also missing a verb to connect "black dog" and "ball." Finally, results (5) and (6) mentioned a wrong object or a wrong action in the caption. We suspect that language modelling is one of the factors that contribute to these errors. Compared with a textual image-captioning model, the Image2Code decoding capabilities are now trained with the discrete speech unit instead of using textual language.

6.5.3 Comparison with Other Systems

Table 6.4: Image2Speech results on Flickr8k dataset in comparison with other systems (phoneme-level evaluation)

No.	Model	Use text?	Output	Synthesized speech		Natural speech		Notes
				BLEU4 ↑	CIDEr ↑	BLEU4 ↑	CIDEr ↑	
	Cascade image → text	✓	text					
(1)	Hout et al., (2020) [116]			-	-	36.1	42.4	Topline for (6)
(2)	Ours (Image2Text, beam 5)			46.21	64.51	46.21	64.51	Topline for (6)
	Cascade image → text → speech	✓	speech					
(3)	Ours (Image2Text→TTS)			43.39	61.09	43.35	60.25	Topline for (6)
	Transfer learning IPA de→en	×	speech					
(4)	Ours (Image2IPA→IPA2Speech)			23.57	21.17	-	-	Baseline for (6)
	Direct image → speech	×	speech					
(5)	Concat ResNet + Tacotron decoder			Did not converge				Baseline for (6)
(6a)	Ours[1] (Image2Speech, greedy)			37.78	49.07	37.73	44.09	Proposed system
(6b)	Ours[1] (Image2Speech, beam 5)			40.09	51.40	41.12	48.22	Proposed system

Table 6.5: Image2Speech results on Flickr8k multi-speaker natural speech dataset (word-level comparison). Proposed approach needs less paired image-speech data compared with previously published results which always need 100% image-speech pairs for training.

Model	{sp}	{img,sp}	BLEU4 ↑	METEOR ↑	CIDEr ↑	ROUGE-L↑
Hsu et al. (2020) [5] (SAT) [2]	-	100%	11.6	14.1	23.2	39.0
Hsu et al. (2020) [5] (SAT-FT) [2]	-	100%	12.5	14.5	24.5	39.1
Ours (Image2Speech, beam 5)	100%	100%	**14.78**	17.40	32.89	45.75
	100%	75%	14.58	16.82	31.07	45.34
	100%	50%	13.93	15.91	28.48	44.21
	100%	25%	9.88	13.43	16.50	41.04

.

[1]Best system from Table 6.2d (synthesized speech) and Table 6.3 (natural speech).

[2]The entire framework was trained with additional datasets: Places dataset [121] for discrete unit learning, and LJSpeech [122] for TTS.

Last, we also compared our best system performance with a previously published system. In addition, we trained a model to be compared as the baseline and topline. First, we trained our image-captioning model with phoneme output, which we call the Image2Text model. This model is similar in principle with Hout et al.'s (2020) previously published results. Our Image2Text model with transformer-based architecture produced better BLEU4 and CIDEr performance compared with Hout et al.'s work (see Table 6.4). Then, we generated the speech utterance of this textual result with a TTS trained using the same dataset. Using the same evaluation method described in Section 6.4.4, this model yields the performance of 43.39 and 43.35 BLEU4 scores for synthesized and natural speech, respectively. We use these scores for the topline of our proposed system because it still uses text as a bridge between the image and speech modality.

Our proposed end-to-end system with an Image2Code beam size of 5 achieved the performance of a 40.09 BLEU4 score for synthesized speech, which is about 3 points away from the topline of image→text→speech pipeline. This means that our proposed discretization agent for a codebook works effectively to replace the text modality. Moreover, the distance to topline is closer with our proposed system on a natural speech dataset, which is about 2 BLEU4 points. Here, our proposed Image2Speech model which does not use any text information during training, outperforms Hout et al.'s previously proposed phoneme-based model by about 5 BLEU4 points in the Flickr8k multi-speaker natural speech dataset.

Then, we also trained a simple end-to-end model with ResNet [4] as an image encoder and the decoder Tacotron [2] to generate speech. We found that the model did not converge and produced unintelligible sounds. We observed that during the teacher forcing of the Tacotron decoder, the Mel-spectrogram input is too long and the speech representation produced is insufficient for use as a query to the attention mechanism to get context from the encoded image representation. This shows that VQ-VAE in our proposed model is crucial for speech feature discretization, where the representation can then be easily associated with the encoded image representation.

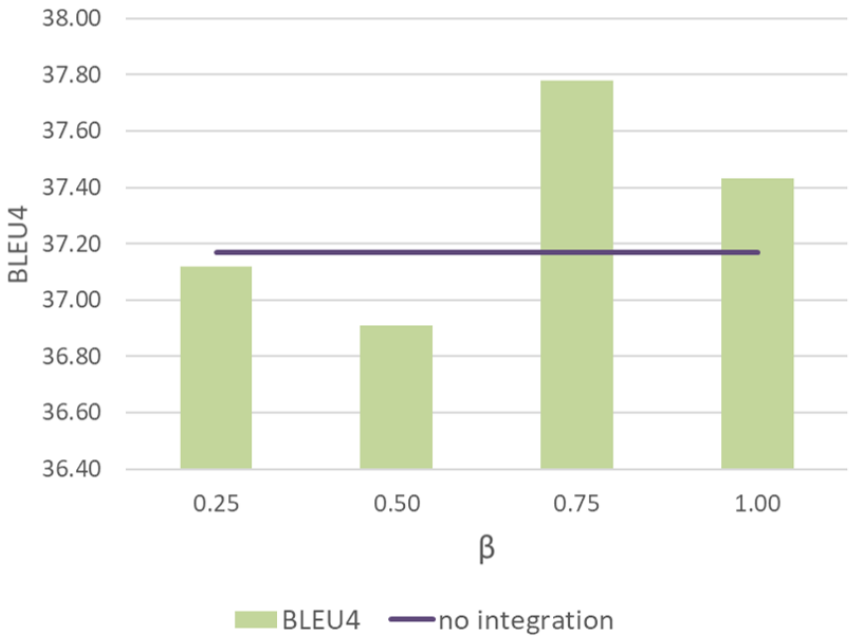

Figure 6.3: Proposed Image2Speech end-to-end integration results compared with cascaded pre-trained model (straight line) on Flickr8k single speaker synthesized dataset (phoneme-level evaluation)

Figure 6.4: Various example results from proposed Image2Speech model trained on multi-speaker natural speech dataset. Caption transcription generated using ASR from the speech caption hypothesis. Images courtesy of Unsplash[1].

[1]https://unsplash.com/; For presentation purposes, we use the example images from the free sources, but they still reflect of what were happening in the Flickr8k test set.

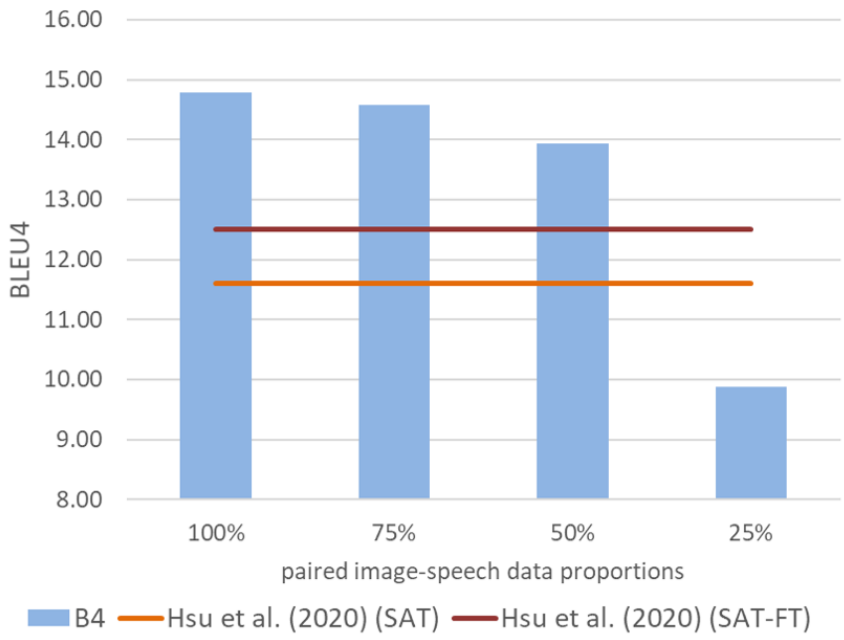

Figure 6.5: Proposed Image2Speech approach compared with Hsu et al. (2020) [5] (red lines) on Flickr8k multi-speaker natural speech dataset (word-level evaluation). Proposed approach (blue bar) can achieve comparable performance even with less than 50% paired image-speech data.

Next, we trained a baseline that is based on transfer learning from German to English in terms of IPA transcription (see Section 6.4.3 for details). As can be seen from Table 6.4, our proposed method performed much better with more than 14 BLEU4 points compared to this baseline. In this baseline, the features of the source language might not be useful for the target language, and some errors might be propagated. This result reveals that our proposed Image2Speech self-supervised discrete speech representation is more effective than such features learned from the cross-lingual approach through transfer-learning. Our proposed approach also can perform in much better quality, even with less data than this baseline, because there is no need for additional data in other languages.

Finally, we also compared our results with Hsu et al. (2020) [5], which also used discrete speech unit representation (see Table 6.5). However, their discrete speech unit is generated from a speech-visual grounding model that needs parallel image-speech data, whereas our approach just needs speech-only data in this step. The proposed approach outperforms their previously published results in all

metrics, regardless of whether fine-tuning of the ResNet model was done as they reported (SAT or SAT-FT). In addition, one of the advantages of the proposed approach is that we do not always need to have the same proportions between image and speech data, in addition to their being paired. Therefore, it is possible to train our model with 100% speech-only data ({sp}) but use less paired image-speech data ({img,sp}), similar to real-life conditions. In both Table 6.5 and Figure 6.5, we demonstrate how our Image2Speech model can outperform the previously published results, even with 50% less paired image-speech data. Moreover, this work's results were achieved without the need for training a separate TTS model, since our proposed approach just needs to use the decoder part of the VQ-VAE to invert the codebook units into speech features.

6.6. Summary

We described our proposed approach in achieving the Image2Speech task without text using the multimodal machine chain framework, inspired by the zero resource speech technology that attempts to provide speech technologies for untranscribed unknown language. Our proposed system uses a pipeline of VQ-VAE, Image2Code, and a Mel2linear inverter. To completely avoid text as a bridge, we used the VQ-VAE codebook to train our image-captioning model, where code sequences can then be inverted into speech features for generating speech. We explored various parameters in the proposed approach and did fine-tuning to achieve end-to-end optimization within the Image2speech pipeline. Our experiment results with a multi-speaker natural speech dataset outperformed previously published work that uses a grounding-based approach, even while using only half of the paired image-speech. This shows the effectiveness of our discrete speech representation in replacing text as the intermediary in the Image2Speech task. Our approach is also more efficient in terms of data size and model size because it accomplishes training with less paired image-speech data than needed by the previously published approach. In addition, the decoder part of our VQ-VAE can also be used during the codebook-to-speech inversion, removing the need to train a separate TTS model.

Chapter 7

Conclusions and Future Directions

In this last chapter, we conclude our thesis and discuss future directions for our proposed framework.

7.1. Problem Reiteration

Human perceives the world with various senses which make their communication multimodal. This perceived information then can be conveyed to other humans in the form of speaking. Denes et al. described how listening and speaking are closely related to each other through a mechanism called the speech chain [1]. In addition, visual modality is also processed together during human communication [10], supplying visual recollection of the viewed objects [9].

Researchers in the speech processing field has been developing machine learning models that mimic this human communication behaviour, in the form of a cross-modal model. However, despite this close relationship in natural human communication, the current research tends to be independently progressing. In addition, it is difficult to introduce a new modality to the system because the more modality we add, the more difficult it is to create the parallel data (i.e. from pair to triplet, quadruplet, and so on). This problem occurs mainly because their approach is mostly focused on supervised learning, which relies heavily on paired data.

7.2. Conclusions

In this section, we review our work from the perspective of theoretical, application, and experimental.

7.2.1 Theoretical Issues

We take advantage of the closeness between various modalities used by humans in communication, to develop a multimodal chain framework that leverages various learning strategies. We generalize the idea of the chain mechanism in the form of a multimodal machine chain (MMC) framework, which aims to enable cross-modal model learning from any kind of data availability. This is possible because we designed a feedback link in the form of reconstruction loss, for any cross-modal operation that has been done inside the chain path.

7.2.2 Application Issues

We showed some proof-of-concept of our proposed MMC framework in various data available for several use cases. First, we showcase our MMC framework capability to enable semi-supervised learning of cross-modal models in various data availability. Second, our MMC framework successfully enables weakly-supervised learning from completely unpaired data in the form of speech-to-text mapping, which previously is not feasible for model training. Then, we successfully develop end-to-end image-to-speech generation by using our framework to learn an optimal speech representation for the task, which uses less paired image-speech data.

7.2.3 Experimental Issues

Each of the applications of our proposed framework enabled a more efficient learning strategy, compared with several existing baselines in each of the tasks. In Chapter 4, we reported a semi-supervised ASR improvement from 36.35% CER to 12.06% CER using unpaired and single-modality data from other unrelated modalities. Then, in Chapter 5, our speech-to-text mapping successfully reaches

14.57 BLEU, while the ASR baseline cannot even converge. Finally, our end-to-end image-to-speech generation can successfully reach 41.12 phoneme BLEU, which is close to the textual topline of 43.35 phoneme BLEU.

7.3. Summary of Contributions

The original contributions of this thesis are listed as follows:

- **A general framework called MMC to enable cross-modal learning by using various levels of supervision (in Chapter 3)**
  Major advantages of the proposed framework:
  - It can be applied to any kind of modalities
  - It can be applied to any kind of cross-modal model
  - It is more effective in particular in low-resource condition
  - It can also be used for untranscribed unknown language, because it can learn an optimal representation when there is none available.
- **Dual-loop multimodal machine chain that combines speech chain and visual chain (in Chapter 4)**
  We showed that it is possible to continue the training of an ASR model, beyond the machine speech chain framework [53]. Our dual-loop mechanism enables ASR model improvement with a single-modality image data, which modality is not even related to ASR (not speech or text). Our experiment result shows that speech processing model performance is improved, while maintaining the performance of other models, in addition to outperforming the label propagation baseline.
- **Multimodal machine chain mechanism that handles multispeaker speech processing (in Chapter 4)**
  We enable multispeaker speech processing by implementing one-shot speaker adaptation in the multimodal machine chain mechanism. The experiment result shows its effectiveness in a multispeaker natural speech dataset.

- **Single-loop multimodal machine chain to show MMC framework usage for multimodal multisource model (in Chapter 4)**
  Using an ImgSp2Txt multimodal multisource model, we showed that it is also possible to use our proposed MMC framework for this kind of model. Our experiment showed that an audiovisual model can be augmented using the single-loop multimodal chain, without any significant performance decrease compared with the dual-loop one.

- **Applications using MMC to enable speech-to-text mapping using unpaired data (in Chapter 5)**
  We investigated a weakly-supervised mapping task to transform unknown untranscribed speech utterances into semantically equivalent texts. The experiment result shows that our proposed partially-aligned Code2Text model and chain augmentation strategy inspired by the MMC framework can successfully perform the mapping even for a cross-lingual application

- **Applications using MMC for end-to-end image-to-speech generation (in Chapter 6)**
  Inspired by the zero resource speech technology that attempts to provide speech technologies for untranscribed unknown language, we use our proposed MMC to attempt image-to-speech generation without text. The key contribution of this approach is the use of transformer-based VQ-VAE to learn the discrete speech representation in a self-supervised manner. Our proposed method outperformed the previously phoneme-based and grounding-based approach, even while using only half of the paired image-speech data.

7.4. Future Directions

Despite all the contributions listed in Section 7.3, we acknowledge that there are still several things that our proposed framework cannot do, such as:

- **Chapter 4: MMCSemiSup**
  Currently, the performance of MMCSemiSup has some percentage gap in terms of CER/WER compared with the supervised topline. To relieve this,

improving the model inside the chain into a more data efficient model can be done. In addition, it is also possible to investigate other possible operation order, considering the levels of supervision. For example, the current implementation of Chapter 4 starts from supervised learning to self-supervised learning. The order can be reversed so that it starts from the least supervision, such as from self-supervised learning to supervised learning.

- **Chapter 5: MMCWeakSup**
  The current implementation of our weakly-supervised Speech2Text mapping still has BLEU score under 20. Although it successfully describe the semantic content of the speech to some extend, the quality of the sentence can be improved, especially in terms of vocabulary modelling, such as described in Table 5.3. Currently, the training relies not only on the chain mechanism and the discrete representation, but also on how we provide the alignment information to connect both partially-aligned speech and text description. Currently, we still rely on an unsupervised aligner to generate the alignment. A better partial alignment modelling strategies can benefit the model performance.

- **Chapter 6: MMCSelfSup**
  The Image2Speech task still produce wrong sentences or unnatural sentences. We analyzed that the error are mostly comes from the difficulties of the model to handle words that are not represented in the image (i.e. not noun), which is also commonly found not only in Image2Speech but also in Image2Text task. We suggest that the use of an adversarial-based image captioning model may relieve this situation. In addition, another order or target of the operation can be investigated too. The model can benefit from the discretization of not only speech, but also image. The discretization of image has been discussed by van den Oord (2018) [57].

- **A modular implementation for easy operation combination**
  Currently, we implement the chaining mechanism using the PyTorch neural network library [123]. When implementing the chain operation, we need to define the forward operation manually for each of the chain paths (refer to Section 3.2). For example, the implementation of $\mathcal{C}_{XYX}$ and $\mathcal{C}_{YXY}$ requires

two different functions. This implementation is of course not ideal for rapid prototyping, especially when we want to reorder the operations based on the levels of supervision (i.e. unsupervised first, then semi-supervised). We suggest developing a modular implementation that considers a modular design that reflects the abstraction of the chain operation.

In general, the remaining gap of performance found in each of our proposed framework implementation is caused by the model component that produces low performance due to various reasons (i.e. not data efficient, bad tuning, etc), and due to operation variations (different order, different target) of the general framework that has not yet been investigated. On the other hand, there are several things that are not yet covered in this thesis, such as:

- **Still need to restart from zero when training inverse model**
  Currently, after finishing the training of $M_{X \rightarrow Y}$ model, we need to re-train the inverse model of $M_{Y \rightarrow X}$ from zero. An example of this is when the ASR model has been fully trained, we still need to train the TTS model from zero, although the acoustic information has been modelled by ASR. It will be ideal if there is an information-sharing mechanism in between the related components of the chain. This will enable us to reduce the number of parameters and training time, and also possibly reducing the number of data needed.

- **A proof-of-concept with other kinds of modality**
  Currently, our proposed MMC framework has shown its capability to handle image, speech, and text modalities. There are still other modalities that have not yet been covered such as videos, sound/audio, and sensor data. It will be good to have a proof-of-concept to showcase that our proposed MMC framework can also be used for such modalities.

- **A more comprehensive multilingual multimodal approach**
  Although we described an attempt for a multilingual approach in Section 5.5.3, our proposed MMC framework has not yet been used for a multilingual approach. We hypothesize that the implementation of a multilingual approach will regard another language as another modality, while image modality can be used as a bridge or conduit.

- **Visual chain implementation lack an auxiliary information modelling**

  The current visual chain implementation has not yet considered auxiliary information passing from the IC to IR/IG. Auxiliary information is the additional factor that is not directly related to the task itself but is important to maintain the consistency of the loop. The example of auxiliary information in the speech chain is the speaker information that is passed from ASR to the TTS so that the reconstructed speech's speakers are still consistent. In the visual chain, the use of auxiliary information can be useful to maintain the consistency of the generated image. For example, when the IC generates a caption "a cat on the table", additional information is needed by IG to reconstruct the image, such as what is the colour of the cat, in what position does the cat stays, and in what direction does the cat looking.

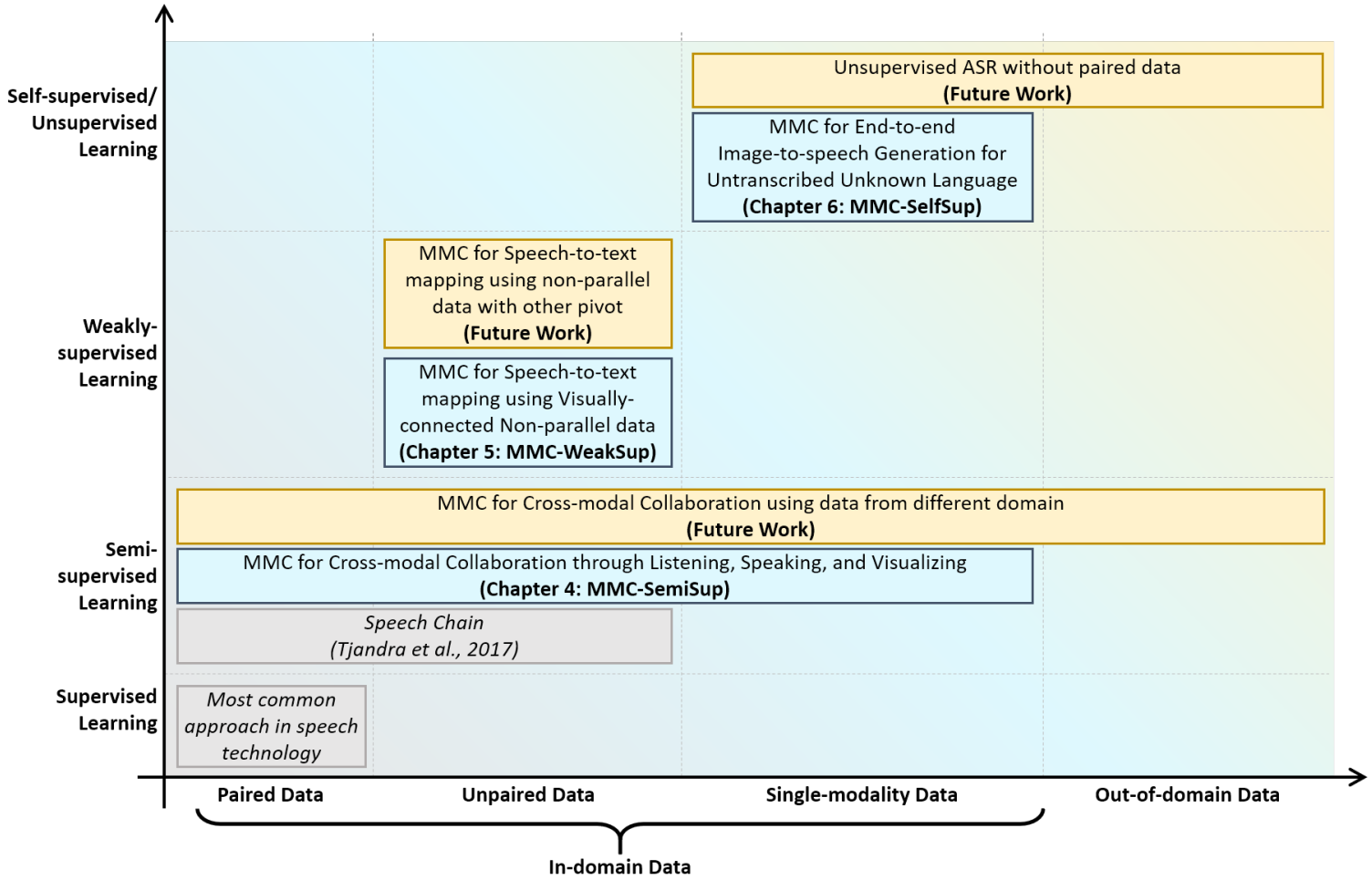

Figure 7.1: Future directions

We discuss the future directions to cover these current limitations of our proposed MMC framework. We redraw the Figure 1.7 in Chapter 1, in the Figure 7.1 with an addition of yellow box. The blue box represents the current scope of our proposed MMC framework, while the yellow box represents possible improvements in terms of data availability and level of supervision. Based on the figure, we describe the future directions as follows:

- **Cross-modal collaboration using data from different domain**
  Our MMC framework now can successfully take advantage of a non-ideal data condition such as unpaired or detached (single-modality). However, all those are based on an assumption that all the data are in-domain. Enabling some domain adaptation methods to use this data will allow the MMC framework to be used for more purposes.

- **Unsupervised ASR without paired data**
  Related to the previous future direction, a speech-to-text mapping model that can learn from purely unpaired data can be further developed into an unsupervised ASR. We define unsupervised ASR here as an ASR model that can be trained without paired data. We suppose that the representation learning from our proposed MMC framework can be also conditioned to some bootstrap information that links speech and text modality.

- **Towards another type of media**
  Currently, we are using a multispeaker speech to represent speech modality, and an image to represent visual modality. There is still various kind of media that are not yet covered in this thesis, such as:

  - noisy speech
  - multilingual speech or code-switching speech
  - sequence of image or video

- **Towards better cross-modal model**
  Going beyond what is shown in the figure, we also consider it important to upgrade the MMC framework implementation with a better cross-modal model. Given data constraint in this thesis, it will be beneficial to have a

cross-modal model that is more data-efficient. For example, Section 5 and 6 have already shown examples on how self-supervised discretization can reduce dimensionality, which enables convergence with fewer data.

- **Enabling information sharing**

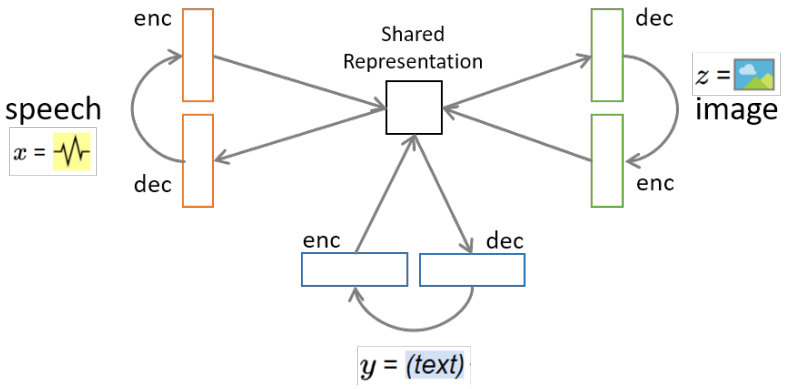

Figure 7.2: Centralized approach with shared multimodal representation.

Currently, each chain component is stand-alone, and there is no information sharing in between them. Therefore, each model needs to be trained from zero. In addition, there is no guarantee that every new information learned in one model can be eventually be propagated in to other models. One practical example is that while both ASR and IC receives different input, but both of them are actually modelling the same language in the decoder part. The combination of those two models, even in the simplest form such as ensembling, has been proven to yield good results in MMCSemiSup2 using ImgSp2Txt model. Therefore, allowing information sharing by implementing a centralized approach will be beneficial for the framework.

- **Multilingual multimodal machine chain**
  For the framework in general, we can regard a different language as a different modality. In addition, image modality can be used as a good bridge in between different languages. Furthermore, there has been some previous works on using the chain mechanism for multilingual purpose. Novitasari et

al. (2020) reported the use of machine speech chain for Indonesian ethnic languages [124]. The ASR and TTS are initially trained using standard Indonesian using supervised training, then the training is continued using the ethnic languages. Therefore, the model can be adapted for those ethnic languages without the use of paired data.

On the other hand, Nakayama et al. (2019) also reported a multilingual machine speech chain for zero-shot code-switching ASR and TTS [125]. Code-switching is defined as when one speaker uses two or more languages interchangeably within a conversation, which then can be classified as a multilingual phenomenon. In their work, the code-switching data is only used for the semi-supervised step, while the supervised step is using monolingual data. Therefore, it decreases the burden of getting parallel code-switching data, which is expensive.

- **Explore various operation order**
  The operation order for the current study is fixed. For example, in MMCSemiSup, the order of operation is first using supervised learning with a small amount of paired data and then continued with semi-supervised learning using unpaired data and unrelated modality data. It will be interesting to also incorporate the self-supervised learning into that training pipeline, given its effectiveness in Chapters 5 and 6 for other tasks.

Appendices

Appendix A

Further Analysis in MMCSemiSup

In this appendix, we provide further analysis on the cross-modal collaboration in MMCSemiSup. First, we do an error analysis to compare the label propagation and the speech chain result in Table 4.5. Then, we discuss the variations of the model such as the size and pretraining factor in the cross-modal collaboration using MMCSemiSup. The result comparison for these are also related to Table 4.4 (for Flickr30k result) and 4.5 (for Flickr8k result).

A.1. Error Analysis on Label Propagation vs MMCSemiSup

The result in Table 4.5 shows that given the same amount of parallel and unpaired data, our MMCSemiSup can improve the baseline while the label propagation method did not. In this section, we do an error analysis to see what kind of error can be fixed and how specifically our proposed method can improve over the label propagation. We focus on the second step, where the initial model training was continued by using the unpaired data $U_{x,y,z}$.

Table A.1: Detailed Comparison between Label Propagation and MMCSemiSup in the 2nd Step of Table 4.5

Factor/Metric	**Baseline**	**Label Propagation**	**MMCSemiSup**
CER	36.35%	39.57%	15.10%
△CER	-	−3.22%	21.25
WER changes	44.04%	47.36%	21.82%
△WER	-	−3.32%	22.22
#Utt Improved	-	1294	3789
#Utt Worsen	-	1841	343
overall	-	−547	3446
BLEU1	62.06	62.39	81.61
BLEU4	38.99	38.53	63.80
METEOR	30.27	30.09	45.24

A.1.1 Quantitative Analysis

We calculated the character and word level error rate (CER/WER), and measured how much improvement does each method contributes in Table A.1. In both CER and WER, the label propagation method failed to continue the training using the unpaired data in a semi-supervised manner. We also compared the number of utterances improved in both methods. We found that in total, there is more utterance get worsen rather than improved in the label propagation, as compared with the MMCSemiSup. In addition, we also calculate the BLEU and METEOR score. Both metrics are improved in the result using MMCSemiSup, which shows that the result is not only improved syntactically but also semantically.

A.1.2 Qualitative Analysis

We also observe the generated transcription, to get some perspective on what kind of errors does our proposed MMCSemiSup relieved, in comparison to the label propagation method. We listed some of the errors, together with the examples.

- **Fixing from a wrong acoustically similar word to the correct word.**

Baseline: a blond woman in a blue shirt **peering** out from a white fence .
LabelProp: a blonde woman in a blue shirt **poses** with a white fence .
MMCSemiSup: a blond woman in a blue shirt appears to wipe for a ride .
Reference: a blond woman in a blue shirt appears to wait for a ride .

Baseline: a boy rides on a trampoline .
LabelProp: a boy rides on a trampoline .
MMCSemiSup: a boy rides on a tire swing .
Reference: a boy rides on a tire swing .

In the first result example, the initial baseline result has tried to find a similar word of "appears" into "peering". Both words are acoustically similar, although it is incorrect. Then, the label propagation method relies on improving the language modelling of the model, which then updates the incorrect word "peering" into "poses", which is still incorrect. Finally, the MMCSemiSup corrects the word into "appears". On the other hand, in the second example, the MMCSemiSup fixes the word "trampoline" into "tire swing", which shows that now the model is more conditioned on the given speech. In both examples, the model trained by MMCSemiSup relies more on the improvement of acoustic modelling, which is enabled by the help of the TTS model in the chain.

- **Reduces word dropping, predicts length better.**

Baseline: a boy is sitting .
LabelProp: a boy sitting in a pool .
MMCSemiSup: a boy sitting in water .
Reference: a boy sitting in water .

Baseline: the man is playing tennis guitar .
LabelProp: the man is playing tennis in the background .
MMCSemiSup: the man is playing tennis against a building .
Reference: the man is playing tennis against the building .

In these two examples, the baseline suffers from a shorter hypothesis compared to the reference due to word droppings. This indicates that the model is not good enough in predicting the length of the transcription, given the speech utterance input. In the first example, label propagation successfully

predicts the correct length but filling it with the incorrect word, while the one using MMCSemiSup can predict the correct word accurately. Similar to the first example, the second example shows how both models predict the same length, but MMCSemiSup gives a better hypothesis than label propagation.

- **Better language modelling: fixes semantics**

Baseline:	a man and woman sitting on a dog .
LabelProp:	a man and woman sitting on a dog .
MMCSemiSup:	a man and woman sitting on a dock .
Reference:	a man and a woman sitting on a dock .

  The word "dog" and "dock" are acoustically similar, but the phrase "sitting on a dog" perhaps is not as common as "sitting on a dock". In this example, our proposed MMCSemiSup fixes the wrong word selection, thanks to better language modelling.

- **Better language modelling: fixes unknown words.**

Baseline:	a young girl sfands with her leaves .
LabelProp:	a young girl floats with her little girl in the leaves .
MMCSemiSup:	a young girl ' s face looking through leaves .
Reference:	a young girl ' s face looking through leaves .
Baseline:	basketball player on the ribber .
LabelProp:	brown and black dog is running through a ring .
MMCSemiSup:	rafting boat on a river .
Reference:	rafting boat on river .

  In both examples, unknown words such as "sfands" and "ribber" are generated. This is because the ASR model is using a character-level granularity, so in addition to word-by-word modelling, the model also needs to do character-by-character modelling to generate a correct word. MMCSemiSup corrects both errors into the correct words.

We found the ASR model trained using MMCSemiSup provides a qualitatively better hypothesis. Several improvements in terms of better acoustic modelling, language modelling, and length prediction can be found in our proposed MMC-Semisup model, as compared with the label propagation.

A.1.3 Continuous Improvement in the Chain Mechanism

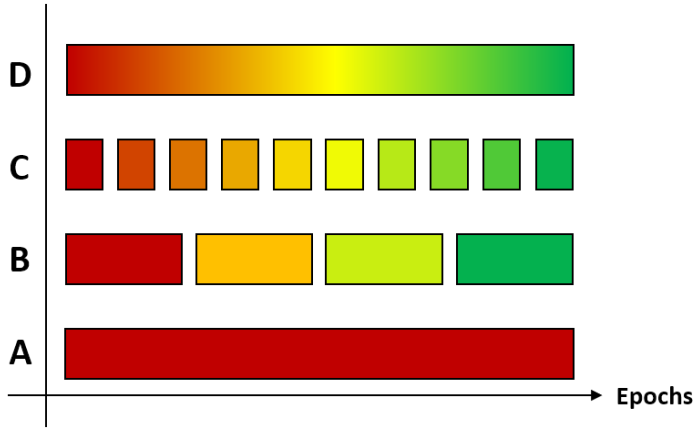

Figure A.1: Model update interval illustration in semi-supervised step. Assume that the model quality improvement is symbolized as a transition from the red to green color. (A) the same model from the first epoch is being used to augment. (B) model is updated in every specified interval. (C) model is updated in a shorter interval. (D) model is updated continuously.

As defined in Section 4.4.5, label propagation generates a pseudo-label from unlabelled data, using the model previously trained using a small set of labelled data. Therefore, with this traditional definition, the pseudo-label itself is generated only at the beginning, using the model before the label propagation starts. The pseudo-label will never be updated, even when the model is getting better due to additional epochs from the label propagation. In Figure A.1, this method is (A), where a model with bad performance (red) produces a bad quality pseudo-label, which are continuously being used throughout the semi-supervised training step.

Then, assume that we set an interval to update the model, whether it is every epoch or every certain number of iterations. The pseudo-label are then generated with an intermediary model, which is assumed to be better than the initial model. As we can see in Figure A.1 B and C, the interval is getting smaller, and the color gradates step by step, representing the quality of the pseudo-label. The training process, as represented by the loss function, are progressing towards convergence, where each steps is better than the previous one. Therefore, the

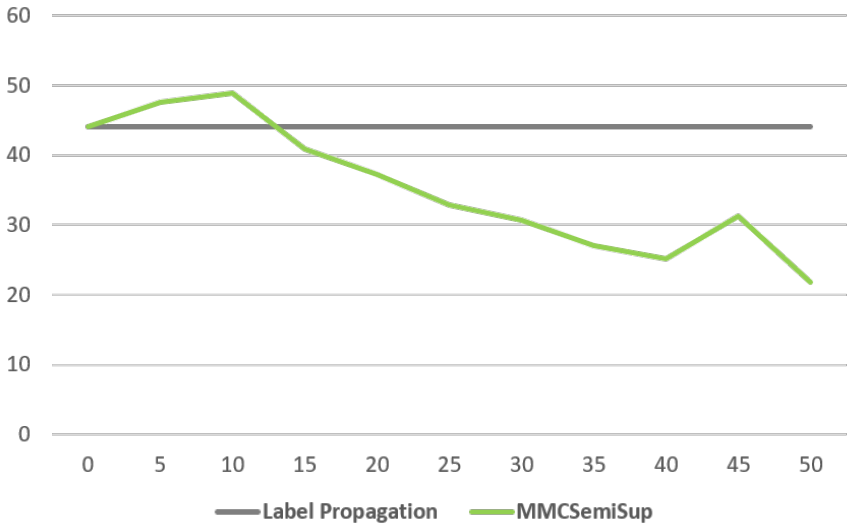

Figure A.2: Quality of the ASR model (in terms of WER) throughout epoch.

more often the pseudo-label is updated, the more effective it is for semi-supervised training. Inspired by this observation, our proposed MMCSemiSup takes this to the next step, by enabling a continuous improvement throughout the semi-supervised training. The pseudo-label is generated when it is needed so that it is generated by the latest version of the model. This ensures that the pseudo-label quality is advancing alongside the model quality.

We show the quality of our ASR and TTS models during the semi-supervised step in every 5 epochs, by running inference using test set (Figure A.2 and A.3). ASR quality here is represented as WER, while TTS quality here is represented as L2 Loss. Using label propagation, the quality of the pseudo-label is always the same from epoch 0. Then, looking at the ASR quality using chain, the increase at the early epochs shows that the pseudo-label itself is not good enough to leverage the training convergence. However, looking at the TTS quality, it is clear that the TTS itself is consistently improving throughout epochs. When the TTS quality is good enough in around epoch 15, the ASR quality can also be directly affected thanks to the continuous improvement. We regard this as a rationale on how our proposed method is more effective than label propagation.

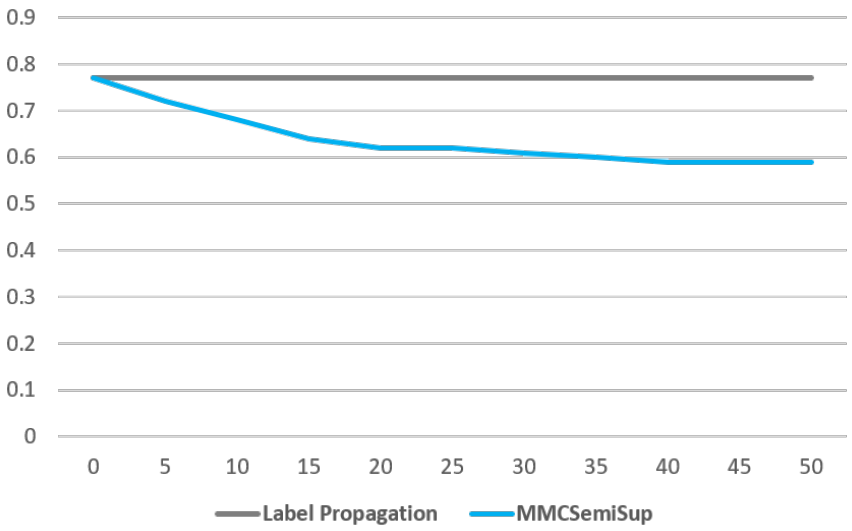

Figure A.3: Quality of the TTS model (in terms of L2 loss) throughout epoch.

Table A.2: Comparison with the current ASR model (refer to Section 4.4.3) with its smaller version.

Parameter	**Current Model**	**Smaller Model**
Bi-LSTM Encoder Size	256 (x2)	128 (x2)
Encoder Depth	3	3
LSTM Decoder Size	512	256
Decoder Depth	1	1
Dropout	0.25	0.25

A.2. Model Size Effect on MMCSemiSup

Table A.2 compares the ASR model used in Chapter 4 (Current Model), with a smaller version of the model which uses about half of the parameter (Smaller Model). The purpose of this additional experiment is to see if the baseline reported in Tables 4.4 and 4.5 are under-trained. We use a model with half of the LSTM size from 512 to 256.

As shown in Table A.3, the smaller model has lower CER in the baseline setting, it is because that model has fewer parameters. However, the performance is almost similar with the bigger model after cross-modal collaboration, with final CER about 2.98% compared with 2.77% with the bigger model. On the other

Table A.3: Comparison of the current and smaller ASR model in the Flickr30k and Flickr8k dataset.

Training	Data Type	Hour(s)	Current				Smaller			
			ASR	**IC**	**TTS**	**IG**	**ASR**	**IC**	**TTS**	**IG**
			CER↓	B4↑	$L2^2$↓	IS↑	CER↓	B4↑	$L2^2$↓	IS↑
Flickr30k	P_{xyz} Multimodal	3.54	21.46	10.55	0.72	4.06	7.93	10.55	0.72	4.06
SingleSpk	$+U_{x,y,z}$ Multimodal	12.55	4.02	10.92	0.49	5.59	4.04	10.92	0.46	5.59
(Semi-Supervised)	$+S_{x,z}$ Sp/Img only	19.97	2.77	11.38	0.43	7.21	2.98	10.58	0.43	6.55
Topline	P_{xyz} Multimodal	51.96	0.68	13.64	0.40	7.57	0.82	13.64	0.40	7.57
Flickr8k	P_{xyz} Multimodal	4.57	36.35	12.75	0.77	5.90	29.90	12.75	0.77	5.90
MultiSpk	$+U_{x,y,z}$ Multimodal	8.57	15.10	13.22	0.59	8.29	20.36	13.22	0.60	8.29
(Semi-Supervised)	$+S_{x,z}$ Sp/Img only	10.70	12.37	13.28	0.56	9.12	17.06	13.83	0.57	9.16
Topline	P_{xyz} Multimodal	34.31	5.76	19.91	0.50	9.66	6.39	19.91	0.50	9.66

hand, the difference is more apparent in the multi-speaker natural speech setting, using the Flickr8k dataset. Initially, the smaller model has a lower CER of 29.90%, as compared to the original model with 36.35% CER. However, after the cross-modal collaboration using a semi-supervised chain mechanism, the bigger model is able to get a better score than the smaller ones. This shows that the bigger model is indeed needed, given the multi-speaker settings which is more difficult.

From a higher perspective, this experiment showcases the confusion in deciding the model size. If the model size is smaller, it will have a better score at the beginning, but less improvement in the next step. On the other hand, a bigger model size gives more room for improvement in the latter step. Therefore, we suggest the use of a bigger model when doing cross-modal collaboration using MMCSemiSup.

A.3. Image Encoder Pretraining Effect on MMCSemiSup

This section compares the effect of using a pretrained or not pretrained ResNet. A pretrained ResNet allows the image to be encoded into a richer high-level feature. On the other hand, we also provide an additional experiment to show that our proposed MMCSemiSup can still be effective even with a not pretrained ResNet.

In the initial supervised step of using P_{xyz} paired dataset, the BLEU score

Table A.4: Comparison of the pretrained and not pretrained ResNet in IC model in the Flickr30k and Flickr8k dataset

Training	Data Type	Hour(s)	Pretrained				Not Pretrained			
			ASR	**IC**	**TTS**	**IG**	**ASR**	**IC**	**TTS**	**IG**
			CER↓	B4↑	$L2^2$↓	IS↑	CER↓	B4↑	$L2^2$↓	IS↑
Flickr30k	P_{xyz} Multimodal	3.54	21.46	10.55	0.72	4.06	21.46	8.38	0.72	4.06
SingleSpk	$+U_{x,y,z}$ Multimodal	12.55	4.02	10.92	0.49	5.59	4.02	8.60	0.49	7.93
(Semi-Supervised)	$+S_{x,z}$ Sp/Img only	19.97	2.77	11.38	0.43	7.21	2.65	10.92	0.43	7.51
Topline	P_{xyz} Multimodal	51.96	0.68	13.64	0.40	7.57	0.68	14.91	0.40	7.57
Flickr8k	P_{xyz} Multimodal	4.57	36.35	12.75	0.77	5.90	36.35	8.40	0.77	5.90
MultiSpk	$+U_{x,y,z}$ Multimodal	8.57	15.10	13.22	0.59	8.29	15.10	9.68	0.59	4.68
(Semi-Supervised)	$+S_{x,z}$ Sp/Img only	10.70	12.37	13.28	0.56	9.12	13.00	12.47	0.57	8.43
Topline	P_{xyz} Multimodal	34.31	5.76	19.91	0.50	9.66	5.76	13.54	0.50	9.66

of IC using a not pretrained ResNet is about 2 points less than the one using a pretrained ResNet. This is because the not pretrained ResNet, given more layers untrained from the beginning, needs more data to train. Then, continuing the training using $U_{x,y,z}$ unpaired data improves the IC model into 9.68 BLEU. Finally, with the cross-modal collaboration using the speech only and the image only dataset, we can get the IC performance with 12.47 BLEU. Even without a pretrained ResNet, the visual chain component can still see some improvement.

Then, looking at the ASR performance, we can see that the CER score can still be improved from 15.10% to 13.00% in the cross-modal collaboration step, even without a pretrained ResNet. This CER is also close to the CER of the one with pretrained ResNet. From this result, we can conclude that our proposed framework can still work even without pretraining, although it is also good to have when it is available.

Appendix B

Discussion on the Tradeoff between Data Size and Quality

B.1. In Cross-modal Collaboration (using MMCSemiSup)

The tradeoff between data size and quality has been discussed in Sections 4.5.5 and 4.5.6.

B.2. In Weakly Supervised Speech2Text Mapping (using MMCWeakSup)

In principle, there are two kinds of data being used to train the framework in Chapter 5:

- Speech only dataset (S_x) to train the VQ-VAE model
- Visually-connected non-parallel speech-text dataset ($U_{x,y}$) to train partially aligned Code2Text and Text2Code

To analyze the tradeoff between data size and quality, we trained another model with the data specifications as described in Table B.1. The 100:100 scenario means that it is using 100% of the speech only dataset S_x and 100% of the $U_{x,y}$

Table B.1: Data availabilities to measure the tradeoff between data size and quality. The scenario 100:100 is similar with the settings described in Section 5.4.1. Percentage reported in data partition size is measured against scenario 100:100.

Scenario	S_x	Ux, y
50:50	6k (50%)	6k (50%)
100:50	12k (100%)	6k (50%)
100:100	12k (100%)	12k (100%)

dataset for training. It is the result reported in Table 5.1 and 5.2, which is using the data specifications described in Section 5.4.1. Then, the 100:50 scenario is decreasing the number of unpaired data $U_{x.y}$ into half of the 100:100, so that it is using 100% of the S_x data, while reducing the use of $U_{x,y}$ data into 50% of the original amount. Finally, 50:50 reduces both single-modality data S_x and unpaired data $U_{x.y}$ into half of the 100:100.

Table B.2: Adapting best Speech2Text model trained on Table 6.2 to the Flickr8k multispeaker natural speech non-parallel dataset

Model	**50:50**			**100:50**			**100:100**		
	Sim%	**BLEU**	**CIDEr**	**Sim%**	**BLEU**	**CIDEr**	**Sim%**	**BLEU**	**CIDEr**
(Synthesized Speech - SingleSpk)									
Code2Text	35.29	12.57	26.28	34.61	12.64	26.37	35.79	15.04	31.66
+Partial Code2Text	37.21	14.94	30.70	31.28	14.86	37.09	40.94	16.80	36.86
+Cycle Augmentation	36.86	15.67	31.75	36.85	14.19	30.02	40.47	17.25	37.52
(Natural Speech - MultiSpk)									
no adaptation	19.25	6.31	9.45	21.58	6.98	11.16	21.31	7.83	11.69
with adaptation	22.85	8.86	15.62	31.76	12.54	24.62	35.35	14.57	29.01

Table B.2 shows the result for each scenario. Comparing 50:50 and 100:50 scenarios, we can find out the effect of adding more single modality data. We found that on the single-speaker data result, the cosine similarity between 50:50 and 100:50 does not change much. We also find that the 50:50 BLEU score with cycle augmentation is higher than 100:50. However, comparing the further adaptation to natural speech, we found that the 50:50 performance is fall behind the 100:50. Therefore, we can conclude that adding more single-modality data is useful when the speech data has a high variability, such as multiple speaker and noisy environment.

Table B.3: Tradeoff between data size and quality. The scenario 100:50 and 100:100 are similar with the settings described in Table 6.5. Percentage reported in data partition size is measured against scenario 100:100.

Scenario	S_x	Pxz	BLEU	CIDEr
50:50	14.5k (50%)	14.5k (50%)	12.34	26.26
100:50	29k (100%)	14.5k (50%)	13.93	28.48
100:100	29k (100%)	29k (100%)	14.78	32.89

Then, we compare the 100:50 and 100:100 experiments to investigate the effect of adding more unpaired data. This allows more training to the partially-aligned Code2Text model. We observe about 2 points of BLEU improvement when adding more unpaired data. We also see some improvement in the multispeaker settings. This shows that adding more unpaired data will generally improve the performance, because it allows more iteration on the Code2Text model.

B.3. In Image2Speech (using MMCSelfSup)

In principle, there are two kinds of data being used to train the framework in Chapter 6:

- Speech only dataset (S_x) to train the VQ-VAE model
- Parallel speech-image dataset (P_{xz}) to train the Image2Code model

Table B.3 shows the tradeoff between adding more single-modality S_x data and paired image-speech P_{xz} data. Similar to the previous section, the 100:100 scenario means that it is using 100% of the speech-only dataset S_x and 100% of the parallel speech-image dataset P_{xz}. By comparing 50:50 and 100:50 scenarios, we can find the effect of adding more single-modality data, which contributes to a VQ-VAE that trained with more data. We found that this increases the BLEU score by 1.59 and CIDEr by 2.22 points. Then, we compare the 100:50 and 100:100 scenarios, which shows the effect of adding more paired data, so that the Image2Code model can be trained with more data. We found that this increases the BLEU score further by 0.85 and CIDEr by 4.41 points.

The first comparison has more BLEU improvement, but the second comparison has more CIDEr improvement. We can conclude that both of them are

equally improved. Consequently, we can regard that both single-modality data and paired image-speech data are both equally important to improve the Image2Speech model using the MMCSelfSup framework.

Appendix C

Discussion on Number of Codes/Clusters in MMCSelfSup

This appendix complements Table 6.2a which compares the number of clusters in MMCSelfSup. We found that the increase in cluster number is not always positively correlated with the end performance because, after convergence, the VQ-VAE model did not use all of the possible codebooks to represent the given speech-caption utterance.

C.1. Number of Codes Effect to the VQ-VAE Losses

In this section, we compare the different number of cluster effects to the overall VQ-VAE loss described in Section 3.2.6 and to the reconstruction loss in particular.

Figure C.1 compares the VQ-VAE loss from different number of clusters. We found that the VQ-VAE with 256 clusters has the lowest loss. Then, we can regard that a cluster number less than 256 is not enough, and higher than 256 is too many. Looking at the figure, we can conclude that when the model has not had enough clusters, the loss will be much higher than when the model has too many clusters.

Then, we can see the reconstruction loss in particular, by looking at Fig-

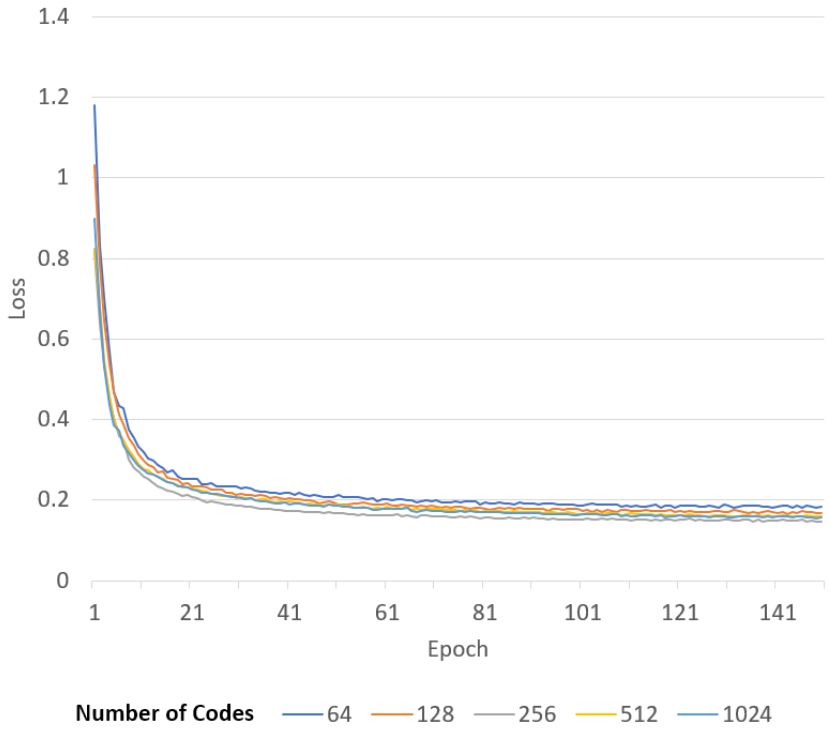

Figure C.1: Comparing the MMCSelfSup VQ-VAE loss of different number of clusters (refer to Table 6.2a). X-axis: loss, Y-axis: epoch.

ure C.2. Compared with the VQ-VAE loss, the effect of the number of codes is more apparent here. The VQ-VAE with 256 clusters has the lowest reconstruction loss, with quite a large margin compared with the other settings. We also can take a similar conclusion, that by assuming 256 is the ideal number of clusters, too many clusters will give lower loss than too few clusters.

C.2. Number of Codes Effect to Codebook Utilization Rate

Figure C.3 compares the utilization rate of VQ-VAE with various number of clusters. After decoding the speech using a trained VQ-VAE model, we get the discrete representation of each utterance. Then, we calculate the unique number of codes that are being used to define the dataset. In this way, we can find out

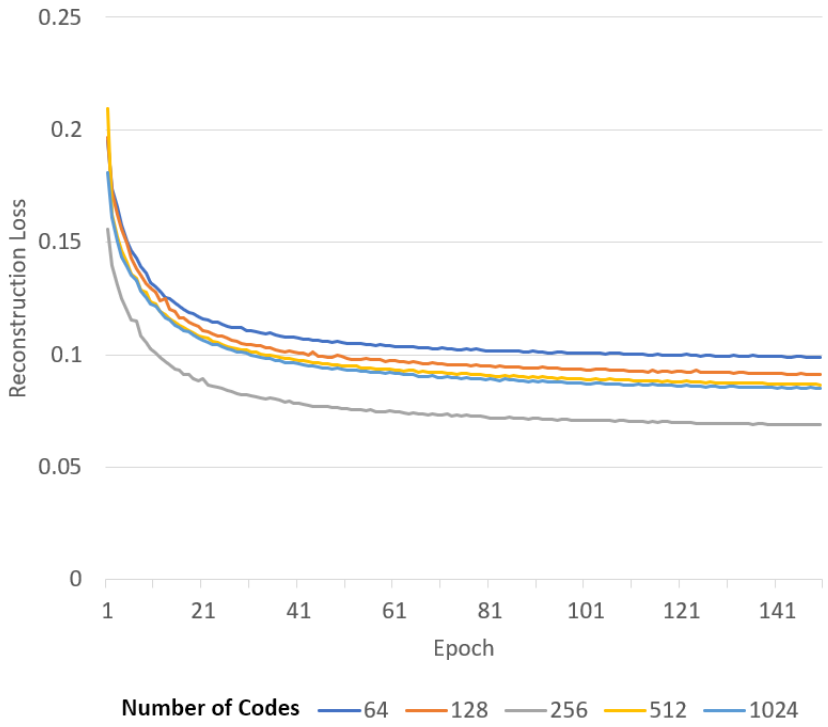

Figure C.2: Comparing the MMCSelfSup reconstruction loss of different number of clusters (refer to Table 6.2a). X-axis: reconstruction loss, Y-axis: epoch.

how many codes are actually being used, given the available number of codes. We call this "codebook utilization rate", which can be used to decide if the number of the codebook is too few or too many for the given dataset.

We found that the number of codes 64 and 128 utilizes 100% of the available code. There are two interpretations of these results. It is either that (a) there is not enough code available, or (b) there is just enough code available. Given that the utilization rate is under 100% in 256 codes setting, we can assume that code number 64 is the case (a), and code number 128 might have a possibility to be the case (b). Considering the losses also in the previous section, we decided to use code number 256 for the next step of the experiment. Nevertheless, the unused code is also not a problem for the Image2Code model. This is because in this case, the Image2Code model vocabulary was build only on the utilized vocabulary (i.e. the one with frequency > 0).

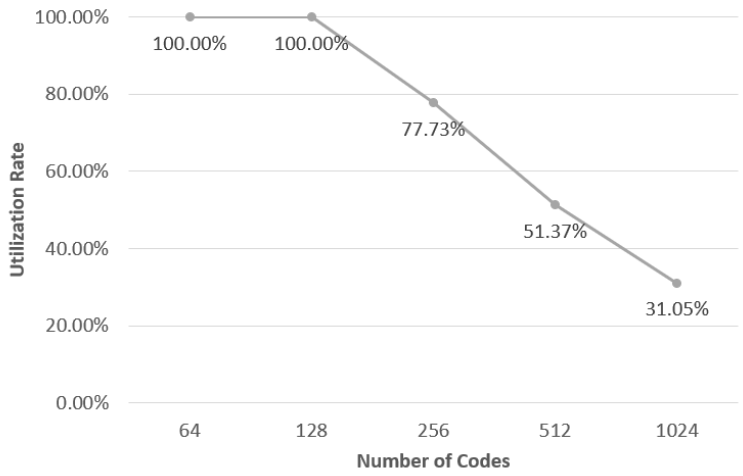

Figure C.3: Codebook utilization rate of different number of clusters (refer to Table 6.2a). X-axis: number of clusters, Y-axis: utilization rate.

C.3. Additional Analysis on Code Sequence Pattern

```
a brown dog is running in the sand .
a brown dog playing in the sand .
a brown dog plays at the beach .
a dog is standing in the sand .
a dog running fast in the sand .

12 19 5 25 14 61 25 60 13 0 36 14 14 13 55 51 10 19 18 44 28
12 34 5 25 14 61 7 32 27 16 11 9 29 20 51 10 19 44 44 0
12 19 5 25 14 20 7 32 56 29 6 20 20 30 32 45 27 44 44 32
12 50 8 47 49 15 61 51 9 9 9 29 20 51 10 19 44 44 0
12 50 43 25 34 5 16 4 55 25 13 58 61 9 29 40 51 14 43 44 44 0
```

Figure C.4: The example where the generated codebook sequence can also consistently represent the overlap in the original speech. (top: speech transcription, bottom: code sequence/discrete representation)

Figure C.4 shows the codebook from speech captions from the same image. As we can see, the same overlap can be found between the speech transcription and the codebook sequence. This consistency of the generated codebook sequence shows how successful our proposed MMCSelfSup was to learn a discrete repre-

sentation. In addition, we found an interesting phenomenon, where the same speech segment can have a slightly different code sequence. For example, the same phrase "a brown dog" has a slightly different code sequences of "12 19 5 25 14 61", "12 **34** 5 25 14 61", and "12 19 5 25 14 **20**" due to the uncertainty in the codebook selection process. This slight difference is the reason we add an end-to-end finetuning, which allows the ambiguity to be handled later in the VQ-VAE decoder. This resulted in a slight improvement as reported in Table 6.2d.

References

[1] P. Denes and E. Pinson. *The Speech Chain*. Anchor books. Worth Publishers, 1993.

[2] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. INTERSPEECH*, pages 4006–4010, 2017.

[3] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Transformer VQ-VAE for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. *Proc. INTERSPEECH*, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, pages 770–778, 2016.

[5] Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In *NeurIPS 2020 Workshop for Self-Supervised Learning for Speech and Audio Processing*, 2020.

[6] William Havard, Laurent Besacier, and Olivier Rosec. SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set. In *Proc. GLU International Workshop on Grounding Language Understanding*, pages 42–46, 2017.

[7] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions:

Constructing a large-scale japanese image caption dataset. In *Proc. ACL*, pages 417–421, Vancouver, Canada, 2017.

[8] National Research Council. *Hearing Loss: Determining Eligibility for Social Security Benefits*. The National Academies Press, Washington, DC, 2004.

[9] Alex Byrne. Recollection, perception, imagination. *Philosophical Studies*, 148(1):15–26, 2010.

[10] Gemma A. Calvert. Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies. *Cerebral Cortex*, 11(12):1110–1123, 12 2001.

[11] W. H. Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26:212–215, 1954.

[12] Dan Jurafsky and James H. Martin. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

[13] James Cooley and John Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

[14] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, 1975.

[15] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[16] F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3):250–256, 1975.

[17] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett. The darpa 1000-word resource management database for continuous speech recognition. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 651–654 vol.1, 1988.

[18] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.

[19] N. Morgan and H. Bourlard. Continuous speech recognition using multilayer perceptrons with hidden markov models. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 413–416 vol.1, 1990.

[20] N. Morgan and H.A. Bourlard. Neural networks for statistical recognition of continuous speech. *Proceedings of the IEEE*, 83(5):742–772, 1995.

[21] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proc. IEEE ICASSP*, pages 4960–4964, 2016.

[22] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. Deep speech 2 : End-to-end speech recognition in english and mandarin. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on*

Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA, 20–22 Jun 2016. PMLR.

[23] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2018.

[24] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall. English conversational telephone speech recognition by humans and machines. In *Proc. INTERSPEECH*, pages 132–136, 2017.

[25] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. The microsoft 2017 conversational speech recognition system. In *Proc. IEEE ICASSP*, pages 5934–5938, 2018.

[26] K. N. Stevens, S. Kasowski, and C. Gunnar M. Fant. An electrical analog of the vocal tract. *The Journal of the Acoustical Society of America*, 25(4):734–742, 1953.

[27] J. L. Flanagan, K. Ishizaka, and K. L. Shipley. Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *The Bell System Technical Journal*, 54(3):485–506, 1975.

[28] Gunnar Fant. Glottal flow: models and interaction. *Journal of Phonetics*, 14(3):393–399, 1986. Voice Acoustics and Dysphonia Gotland, Sweden, August 1985.

[29] N. Dixon and H. Maxey. Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio and Electroacoustics*, 16(1):40–50, 1968.

[30] J. Olive. Rule synthesis of speech from dyadic units. In *ICASSP '77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 568–570, 1977.

[31] Jonathan Allen, M. Sharon Hunnicutt, Dennis H. Klatt, Robert C. Armstrong, and David B. Pisoni. *From Text to Speech: The MITalk System*. Cambridge University Press, USA, 1987.

[32] D. Klatt. The klattalk text-to-speech conversion system. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 1589–1592, 1982.

[33] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlche Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[34] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech, 2021.

[35] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J. Weiss, and Yonghui Wu. Parallel tacotron: Non-autoregressive and controllable TTS. *CoRR*, abs/2010.11439, 2020.

[36] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Jia Ye, R. J. Skerry-Ryan, and Yonghui Wu. Parallel tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling. *CoRR*, abs/2103.14574, 2021.

[37] Stavros Petridis, Yujiang Wang, Zuwei Li, and Maja Pantic. End-to-end audiovisual fusion with LSTMs. In *Proc. of AVSP*, pages 36–40, 2017.

[38] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *Proc. of IEEE transactions on pattern analysis and machine intelligence*, 2018.

[39] Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, 2017.

[40] Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig, and Satoshi Nakamura. Nict-naist system for wmt17 multimodal translation task. In *WMT*, 2017.

[41] Robert Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive visual text-to-speech using active appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[42] Jonathan Parker, Ranniery Maia, Yannis Stylianou, and Roberto Cipolla. Expressive visual text to speech and expression adaptation using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4920–4924, 2017.

[43] R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.

[44] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. In *Tech. Rep.*, 2002.

[45] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[46] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Proc. of NIPS*, pages 820–828, 2016.

[47] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proc. of ICML*, pages 1857–1865, 2017.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of IEEE ICCV*, 2017.

[49] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proc. of ICCV*, pages 2868–2876, 2017.

[50] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking: Speech chain by deep learning. In *Proc. IEEE ASRU*, pages 301–308, 2017.

[51] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Machine speech chain with one-shot speaker adaptation. In *Proc. of INTERSPEECH*, pages 887–891, 2018.

[52] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. End-to-end feedback loss in speech chain framework via straight-through estimator. In *Proc. of IEEE ICASSP*, pages 6281–6285, 2019.

[53] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989, 2020.

[54] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux. The zero resource speech challenge 2017. In *Proc. IEEE ASRU*, pages 323–330, 2017.

[55] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[56] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, April 1984.

[57] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Proc. NIPS*, volume 30, pages 6306–6315. Curran Associates, Inc., 2017.

[58] Yunlong Jiao, Adam Gabrys, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa, and Viacheslav Klimkov. Universal neural vocoding with parallel wavenet, 2021.

[59] M. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *Aiche Journal*, 37:233–243, 1991.

[60] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proc. ICLR 2014*, 2014.

[61] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*, pages 2048–2057, 2015.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NIPS*, pages 5998–6008. 2017.

[63] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proc. of CVPR*, 2018.

[64] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, December 2002.

[65] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 08 2017.

[66] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proc. of IEEE CVPR*, pages 3444–3453, 2017.

[67] F. Sun, D. Harwath, and J. Glass. Look, listen, and decode: Multimodal speech recognition with images. In *Proc. of IEEE SLT*, pages 573–578, Dec 2016.

[68] Shigeki Karita, Shinji Watanabe, Tomoharu Iwata, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6166–6170, 2019.

[69] Qiuyuan Huang, Pengchuan Zhang, Dapeng Oliver Wu, and Lei Zhang. Turbo learning for captionbot and drawingbot. *CoRR*, abs/1805.08170, 2018.

[70] 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)*, pages 2641–2649, Dec 2015.

[71] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proc. NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[72] Johanes Effendi, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking and visualizing: Improving ASR through multimodal chain. In *Proc. IEEE ASRU*, pages 471–478, 2019.

[73] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[74] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *Proc. of IEEE ICASSP*, pages 4835–4839, 2017.

[75] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[76] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318, 2002.

[77] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the advances in neural information processing systems (NIPS)*, pages 2234–2242, 2016.

[78] Douglas B Paul and Janet M Baker. The design for the wall street journal-based CSR corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

[79] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philémon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949, 2016.

[80] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Multi-scale alignment and contextual history for attention mechanism in sequence-to-sequence model. In *Proc. of IEEE SLT*, pages 648–655, 2018.

[81] Armand Vilalta, Dario Garcia-Gasulla, Ferran Parés, Eduard Ayguadé, Jesus Labarta, E Ulises Moya-Sánchez, and Ulises Cortés. Studying the impact of the full-network embedding on multimodal pipelines. *Semantic Web*, (Preprint):1–15.

[82] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. CycleGAN-vc2: Improved cycleGAN-based non-parallel voice conversion. In *Proc. IEEE ICASSP*, pages 6820–6824, 2019.

[83] Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda. The NU non-parallel voice conversion system for the voice conversion challenge 2018. In *Proc. Odyssey The Speaker and Language Recognition Workshop*, pages 211–218, 2018.

[84] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *Proc. IEEE SLT*, pages 266–273, 2018.

[85] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda. Non-parallel voice conversion with cyclic variational autoencoder. *arXiv preprint arXiv:1907.10185*, 2019.

[86] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

[87] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.

[88] Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proc. ACL*, pages 3083–3089, 2019.

[89] Hongxiao Bai, Mingxuan Wang, Hai Zhao, and Lei Li. Unsupervised neural machine translation with indirect supervision. *arXiv preprint arXiv:2004.03137*, 2020.

[90] L. Sarı, S. Thomas, and M. Hasegawa-Johnson. Training spoken language understanding systems with non-parallel speech and text. In *Proc. ICASSP*, pages 8109–8113, 2020.

[91] Yining Wang, Yang Zhao, Jiajun Zhang, Chengqing Zong, and Zhengshan Xue. Towards neural machine translation with partially aligned corpora. *Proc. IJCNLP*, 2017.

[92] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *Proc. IEEE ASRU*, pages 237–244, 2015.

[93] Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. Leveraging neural caption translation with visually grounded para-

phrase augmentation. *IEICE Transactions on Information and Systems*, 103(3):674–683, 2020.

[94] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, pages 740–755, 2014.

[95] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. Librosa: Audio and music signal analysis in python. pages 18–24, 2015.

[96] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL '13)*, Atlanta, GA, USA, 2013.

[97] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proc. IEEE CVPR*, pages 4566–4575, 2015.

[98] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proc. EMNLP*, 2019.

[99] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[100] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.

[101] Gilles Adda, Sebastian Stueker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel,

Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the unwritten language barrier: The BULB project. In *Proc. SLTU*, pages 8–14, 2016.

[102] Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The zero resource speech challenge 2019: TTS without T. In *Proc. INTERSPEECH*, pages 1088–1092, 2019.

[103] Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The zero resource speech challenge 2020: Discovering discrete subword and word units. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Proc. INTERSPEECH*, pages 4831–4835. ISCA, 2020.

[104] Lucas Ondel, Lukaš Burget, and Jan Cernocky. Variational inference for acoustic unit discovery. *Proc. SLTU*, 81:80 – 86, 2016.

[105] Michael Heck, Sakriani Sakti, and Satoshi Nakamura. Unsupervised linear discriminant analysis for supporting DPGMM clustering in the zero resource scenario. *Proc. SLTU*, 81:73 – 79, 2016.

[106] O. Scharenborg, L. Besacier, A. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stueker, P. Godard, M. Mueller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, and E. Dupoux. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "speaking rosetta" JSALT 2017 workshop. In *Proc. IEEE ICASSP*, pages 4979–4983, 2018.

[107] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and S. Nakamura. VQVAE unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019. *Proc. INTERSPEECH*, 2019.

[108] Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux.

Unsupervised learning of acoustic subword units. In *Proc. ACL*, pages 165–168, 2008.

[109] L. Badino, C. Canevari, L. Fadiga, and G. Metta. An auto-encoder based approach to unsupervised learning of subword units. In *Proc. IEEE ICASSP*, pages 7634–7638, 2014.

[110] M. Huijbregts, M. McLaren, and D. van Leeuwen. Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In *Proc. IEEE ICASSP*, pages 4436–4439, 2011.

[111] Y. Zhang and J. R. Glass. Towards multi-speaker unsupervised speech pattern discovery. In *Proc. IEEE ICASSP*, pages 4366–4369, 2010.

[112] A. Jansen and B. Van Durme. Efficient spoken term discovery using randomized algorithms. In *Proc. IEEE ASRU*, pages 401–406, 2011.

[113] O. Räsänen, G. Doyle, and Michael C. Frank. Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Proc. INTERSPEECH*, pages 3204–3208, 2015.

[114] A. Tjandra, S. Sakti, and S. Nakamura. Speech-to-speech translation between untranscribed unknown languages. In *Proc. IEEE ASRU*, pages 593–600, 2019.

[115] M. Hasegawa-Johnson, A. Black, Lucas Ondel, O. Scharenborg, and Francesco Ciannella. Image2speech : Automatically generating audio descriptions of images. In *Proc. ICNLSSP*, 2017.

[116] Justin van der Hout, Zoltan D'Haese, Mark Hasegawa-Johnson, and Odette Scharenborg. Evaluating Automatically Generated Phoneme Captions for Images. In *Proc. INTERSPEECH*, pages 2317–2321, 2020.

[117] S. Ma, D. Mcduff, and Y. Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proc. ICCV*, pages 7597–7606, 2019.

[118] D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, 2016.

[119] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[120] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

[121] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Proc. NIPS*, volume 29, pages 1858–1866, 2016.

[122] Keith Ito and Linda Johnson. The LJ speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[123] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[124] Sashi Novitasari, Andros Tjandra, S. Sakti, and Satoshi Nakamura. Cross-lingual machine speech chain for javanese, sundanese, balinese, and bataks speech recognition and synthesis. In *Proc. SLTU*, 2020.

[125] Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Zero-shot code-switching asr and tts with multilingual machine speech chain. In *Proc. ASRU*, pages 964–971, 2019.

List of publications

Publications

Journal Paper (peer-reviewed)

1. *Multimodal Chain: Cross-Modal Collaboration Through Listening, Speaking, and Visualizing*
   Johanes Effendi, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura
   IEEE Access, Vol. 9, pp. 70286-70299, May 2021
   (Related to Chapter 4)

2. *End-to-End Image-to-Speech Generation for Untranscribed Unknown Languages*
   Johanes Effendi, Sakriani Sakti, Satoshi Nakamura
   IEEE Access, Vol. 9, pp. 55144-55154, April 2021
   (Related to Chapter 6)

3. *Leveraging Neural Caption Translation with Visually Grounded Paraphrase Augmentation*
   Johanes Effendi, Katsuhito Sudoh, Sakriani Sakti, Satoshi Nakamura
   IEICE Transactions on Information and Systems, Vol. E103-D, Issue 3, pp. 674-683, March 2020
   (Related to Chapter 5)

International Conference Paper (peer-reviewed)

1. *Weakly-supervised Speech-to-text Mapping with Visually Connected Non-parallel Speech-text Data using Cyclic Partially-aligned Transformer*

Johanes Effendi, Sakriani Sakti, Satoshi Nakamura
Proceedings of INTERSPEECH 2021, September 2021 (to appear)
(Related to Chapter 5)

2. *Augmenting Images for ASR and TTS through Single-loop and Dual-loop Multimodal Chain Framework*
Johanes Effendi, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura
Proceedings of INTERSPEECH 2020, pp. 4901-4905, October 2020
(Related to Chapter 4)

3. *Neural Speech Completion*
Kazuki Tsunematsu, **Johanes Effendi**, Sakriani Sakti, and Satoshi Nakamura
Proceedings of INTERSPEECH 2020, pp. 2742-2746, October 2020

4. *Listening while Speaking: Improving ASR through Multimodal Chain*
Johanes Effendi, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura
Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, pp. 471-478, December 2019
(Related to Chapter 4)

5. *Multi-paraphrase Augmentation to Leverage Neural Caption Translation*
Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, Satoshi Nakamura
Proceedings of the 15th International Workshop on Spoken Language Translation, pp. 181-188, October 2018

6. *Corpus Construction and Semantic Analysis of Indonesian Image Description*
Khumaisa Nur'aini, **Johanes Effendi**, Sakriani Sakti, Mirna Adriani and Satoshi Nakamura
Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) 2018, pp. 42-46, August 2018

7. *Creation of a Multi-paraphrase Corpus based on Various Elementary Operations*
Johanes Effendi, Sakriani Sakti, Satoshi Nakamura
Proceedings of Oriental COCOSDA 2017, pp. 177-182, November 2017

8. *A two-stage emotion detection on Indonesian tweets*
   Johanes Effendi The, Alfan Farizki Wicaksono, Mirna Adriani
   Proceedings of International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 143-146, October 2015

Domestic Conference Paper

1. *Improving ASR with Multimodal Machine Chain*
   Johanes Effendi, Andros Tjandra, Sakriani Sakti, Satoshi Nakamura
   Spring Meeting of the Acoustical Society of Japan (ASJ), March 2021
   (Related to Chapter 4)

2. *From Speech Chain to Multimodal Chain: Leveraging Cross-modal Data Augmentation for Semi-supervised Learning*
   Johanes Effendi, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura
   26th Conference on Natural Language Processing (NLP2020), March 2020
   (Related to Chapter 4)

3. *Enhancing Neural Machine Translation with Image-based Paraphrase Augmentation*
   Johanes Effendi, Sakriani Sakti, Katsuhito Sudoh, Satoshi Nakamura
   25th Conference on Natural Language Processing (NLP2019), March 2019

4. *Visual Description Paraphrase Corpus Creation with Various Elementary Operations*
   Johanes Effendi, Sakriani Sakti, Satoshi Nakamura
   Autumn Meeting of the Acoustical Society of Japan (ASJ), September 2018