Pre-training Approaches for Voice Conversion to Address Data Scarcity and Their Applications to Ground-Truth-Free Tasks

Department of Intelligent Systems
Graduate School of Informatics
Nagoya University

Wen-Chin Huang

Contents

Abstract

Voice conversion (VC) refers to the task of converting one type of speech to another without changing the linguistic contents and has the potential to be employed in medical, business, and entertainment applications. Most pioneering works in VC first require the collection of a parallel dataset, which refers to a set of utterances from the source and the target with the same contents. Then, a frame-based model is trained, which tries to find a mapping for each source speech frame.

As VC techniques evolved, two mainstream approaches were developed to solve the shortcomings of the above-mentioned method. The first type is sequence-to-sequence (seq2seq) modeling, which is designed to tackle problems where the lengths of the source and target sequences differ. When applied to VC, seq2seq models excel in modeling prosody, which correlates to speaker identity performance. The second line of work attempts to make use of non-parallel datasets. A representative approach is the recognition-synthesis (rec-syn) framework, which decomposes the VC function into a recognizer that extracts linguistic contents, followed by a synthesizer that injects the desired target information to generate the converted speech.

This thesis contributes to further addressing the data scarcity issues that hide in the advancement as mentioned above in VC research. The main concept is to apply pre-training, which is a prevailing paradigm in the modern machine learning era. The first problem is the high dataset size requirement of seq2seq VC models, owing to

the complexity of learning such a complex mapping function. A novel pre-training framework based on text-to-speech (TTS) and automatic speech recognition (ASR) was proposed, which was inspired by the information perspective of the three tasks. The core idea is to transfer the linguistically rich hidden representation space in TTS and ASR to VC. The main result is the availability to use only five minutes of parallel data to train a seq2seq VC model.

The second question is whether more data can benefit the recognizer in rec-syn-based VC. Specifically, the potential of applying self-supervised speech representations (S3Rs) to rec-syn-based VC was studied. Given the supremacy of self-supervised learning (SSL) in research fields such as computer vision and natural language processing, it is highly expected that S3Rs can benefit rec-syn-based VC. The main result is a collection of scientific activities, where the core is an open-sourced toolkit named S3PRL-VC that supports a unified experimental environment, including the dataset, tasks, model architecture, and evaluation protocols. A large-scale, systematical study of S3R-based VC is carried out using the toolkit. It is expected that both VC and S3R researchers can gain fruitful insights from the results: for the S3R community, using VC as the downstream task enables the investigation of the S3R model's ability to disentangle speaker and content information; for the VC community, this is by far the largest unified comparative study of S3R-based VC, which could serve as a guide for researchers who wish to continue on this direction.

Finally, the focus is turned to solving a certain type of VC application where the ground truth training target is unavailable. For instance, to enhance the naturalness of dysarthric speech, which is generated by patients suffering from neural diseases, one might wish to collect the normal version of the patient to train a VC model, which is impossible. Similarly, collecting native speech from a non-native speaker is

crucial in training a foreign accent conversion (FAC) model, which is also impossible. A cascade approach that combines seq2seq and rec-syn-based VC models was first proposed to tackle this issue. On the dysarthric-to-normal VC task, it was shown that the naturalness could be improved while the speaker identity preservation needed to be improved. Similarly, on the normal-to-dysarthric VC task, the severity could be simulated while the speaker identity was not completely maintained.

On the task of FAC, along with the above-mentioned cascade method, two other approaches that also utilized the combination of a seq2seq VC model and a rec-syn-based VC were systematically evaluated. Experimental evaluation results showed that the three compared methods had their pros and cons, all of which show the potential of applying these methods to solve these ground-truth-free VC tasks. However, it was also revealed that due to the ground-truth-free property, when evaluating the VC systems of these tasks, the evaluation protocol needed to be re-designed to make the results more trustworthy.

To summarize, the idea of pre-training was applied to tackle the data scarcity problems in current mainstream VC approaches. The experimental results as well as the discussions and insights advanced the research field, and have opened up new directions for future researchers.

1 Introduction

1.1 General background

Voice conversion (VC) aims to convert the speech from a source to that of a target without changing the linguistic content. Speaker conversion, the process of converting speech from a source speaker to a target speaker, is one of the most known VC applications. Despite the many VC applications beyond speaker conversion, to put it more broadly, all VC applications can be seen as a means to an ultimate goal: *unconstrained speech communication*.

Figure 1.1 illustrates how the physical condition of the human body limits the production of human speech [1]. For instance, damaged speech organs cause severe vocal disorders, and the deficient control of the organs can also end up in an accented voice, while the intention is to speak a foreign language natively. What if one can recover disabled functions, or even augment our body to enhance communication abilities? By building VC systems such as speaking aid devices to convert electrolaryngeal (EL) speech to the original voices of patients with vocal cord damage [2–4], or converting accented speech of foreigners into native speech [5, 6], speech communication can be made beyond physical constraints.

Data scarcity has been an all-time challenge for VC, and the problem has become even more severe in the deep learning era. From early statistical methods to recent deep neural network (DNN)-based models, almost all approaches to VC are data-

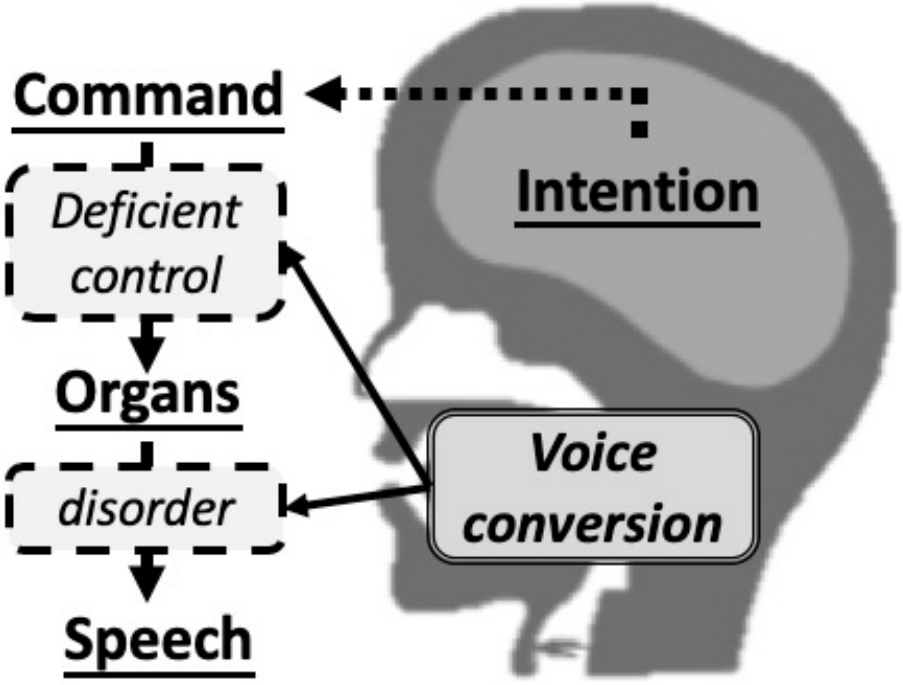


Figure 1.1: *Illustration of the limits of speech communication, and how voice conversion can break the barrier.*

driven. However, it can be assumed that users of VC applications are hesitant to record numerous speech prompts with specific contents because of the laborious process. Although there has not been a standard for the amount of data that a user is usually willing to record, looking back on the past voice conversion challenges (VCCs) [7–9], it could be observed that only several minutes of speech data was assigned to each source or target speaker. On the other hand, datasets for other speech synthesis tasks like text-to-speech (TTS) usually contain hours of data, such as LJSpeech (24 hrs) [10] and VCTK (44 hrs) [11]. Thus, the focus of this thesis will be on addressing data scarcity in VC. To address such an issue, one must first understand current mainstream frameworks and approaches to VC.

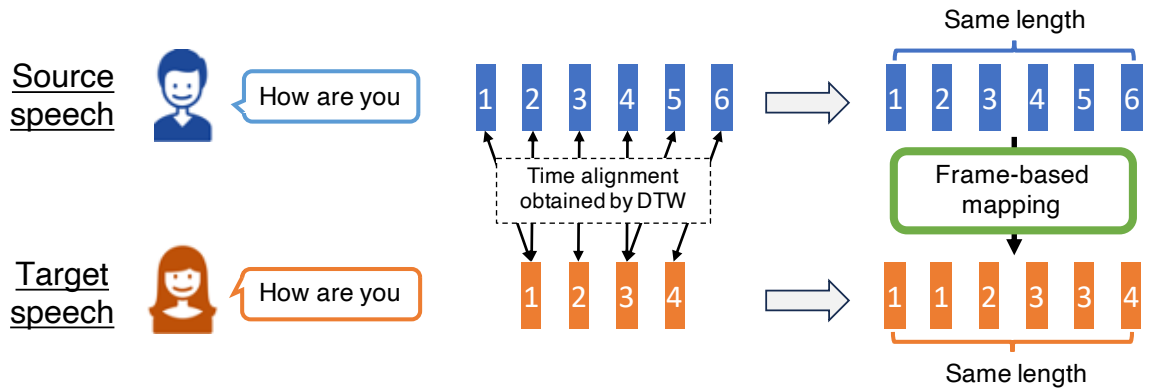


Figure 1.2: *Illustration of the training process of a voice conversion framework based on parallel training data.*

1.2 Thesis Scope

The earliest VC research dates back to the 1980s [12], and has been an active research field for over four decades [13, 14]. If we categorize VC research concerning the data and granularity point of view, most VC research before the 2010s adopted *frame-based* modeling with a *parallel corpus*. A parallel corpus refers to a set of utterance pairs from the source and target with identical content (or prompt), and methods or models that require a parallel corpus to work are often referred to as *parallel VC*. As previously stated, it is often assumed that a parallel corpus as small as several minutes is available [7–9] due to the laborious dataset collection process.

Figure 1.2 shows the training process of a typical VC framework with a parallel corpus. The importance of using a *parallel corpus* is that it is crucial to solving the time alignment problem. Usually, the source and target utterances will have different lengths since the two speakers speak at different rates. To address such a temporal difference, one can assume that the same contents spoken by different speakers are close under a certain distance measure in a particular acoustic feature space. Then, a

time alignment approach can be used to align the source and target acoustic features, where the dynamic time warping (DTW) algorithm is a common choice. The final product will then be aligned with source and target acoustic feature sequences of the same length. Finally, a *frame-based* model can be learned using the paired acoustic frames.

This frame-based, parallel VC framework results in three major problems. In the following subsections, we describe the difficulties and the corresponding solutions which were mostly developed in the late 2010s.

1.2.1 Problem 1: performance

A straightforward pitfall of frame-based modeling is that the input and output acoustic feature sequences will always be of the same length, which means that the temporal structure is left unchanged. This can lead to poor prosody and speaking style conversion, resulting in degraded speech quality and similarity.

To address this problem, researchers have applied sequence-to-sequence (seq2seq) models to VC [15]. Seq2seq modeling refers to an end-to-end DNN-based approach that learns the mapping of the input and output sequences with different lengths, without making assumptions about the structures (as a contrary example, connectionist temporal classification (CTC) assumes a monotonic alignment between the inputs and the outputs). It was shown that seq2seq VC can outperform frame-based models [15,16] regarding both naturalness and conversion similarity.

However, as seq2seq models are based on DNNs, they generally require more data to be successfully trained. As seq2seq VC models still require the availability of a parallel corpus, such a requirement becomes an even bigger problem. In most seq2seq VC papers [15–18], it was reported that roughly one hour of parallel data was used,

Figure 1.3: *The recognition-synthesis voice conversion framework.*

which was much more than the several minutes of data requirement mentioned in Section 1.1. This raises the following research question: **how can we reduce the parallel dataset requirement of seq2seq VC models?**

1.2.2 Problem 2: data requirement

In Section 1.2.1, the research question was how to reduce the parallel dataset size. What if we do not want to use parallel datasets at all? This leads to the development of VC methods that do not use a parallel corpus, and such methods (or models) are often referred to as *non-parallel VC*. Note that while some define a non-parallel corpus to be two sets of utterances from the source and target speakers respectively, as using more data is generally encouraged in modern machine learning, from the practical point of view it is no longer necessary to use only utterances from the source and target speakers.

Among the many methods for non-parallel VC, *recognition-synthesis* (rec-syn) based VC is one of the most adopted approaches. Figure 1.3 illustrates the rec-syn VC framework. The rec-syn VC framework is inspired by the most basic definition of VC, i.e. the preservation of the linguistic contents in the source speech. First, a *recognizer* extracts the so-called *intermediate representation* from the source speech. Then, a *synthesizer* injects information of the target back to generate the final converted speech. Theoretically, the performance of such a rec-syn VC framework highly relies

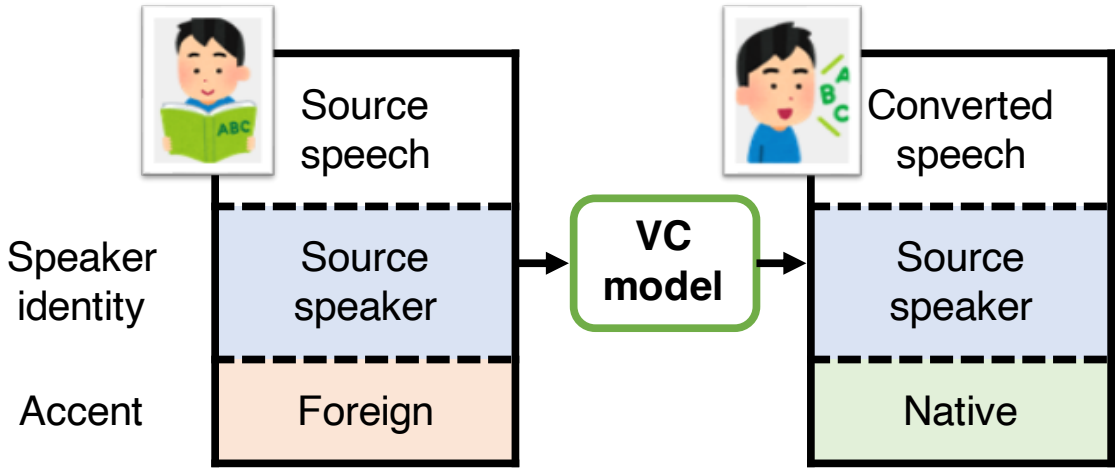


Figure 1.4: *Illustration of the task of foreign accent conversion with identity preservation.*

on the *disentanglement ability* of the recognizer: if the recognizer fails to extract pure linguistic information such that a certain amount of source information remains, then the synthesizer will have a hard time generating converted speech with the desired target attributes since source information resides. Thus, the better the recognizer can disentangle (or remove) unwanted information, the better the performance will be. As both the recognizer and synthesizer are often DNN-based models, more data should ensure better quality. The research question is therefore raised: **how can we utilize more data to improve the performance of the overall performance?**

1.2.3 Problem 3: Unavailability of training target

Certain VC applications aim at *changing certain speech attributes but not the speaker identity*, which makes collecting the training target unavailable. For instance, an application we will be investigating in later chapters is *foreign accent conversion (FAC)*, which is illustrated in Figure 1.4.

In this thesis, we refer FAC to the task of "de-accenting", and readers should avoid confusing this task to a broader term, "accent conversion". While in the literature, accent conversion can be referred to as the conversion between different accents or from native to accented speech, here FAC refers to converting from non-native to native speech. Specifically, given an accented speech utterance spoken by a non-native source speaker, FAC aims to (1) generate a native-sounding version, while (2) preserving the same speaker identity as the source speaker.

The second requirement makes FAC difficult because it is simply impossible to collect native speech from a non-native speaker. One may try to apply Seq2seq VC models described in Section 1.2.1 are not applicable since parallel corpus will not be available. Non-parallel VC methods described in Section 1.2.2, however, will also not be a suitable solution. An important drawback of most current non-parallel VC methods is that they are mostly frame-based models. As stated in Section 1.2.1, frame-based models are poor at converting prosody, which plays an important role in many applications, including FAC.

The research question here is: **how can we approach these VC applications where the training target is unavailable?**

1.2.4 Solution: Pre-training

A common approach to address data scarcity is *pre-training*. As opposed to *training from scratch* (i.e. from random initialization), model pre-training is a paradigm for training DNN-based models using a set of model parameters learned from a dataset of significantly larger scale, regardless of whether the domain of the pre-training dataset or task [19]. The idea is that feature representation learned on a pre-training task contains useful information that is transferable to the target task at hand.

Pre-training is a common training paradigm in the field of computer vision. In the early 2010s, a *supervised* pre-training task (e.g. ImageNet classification [20–22]) was often adopted to enhance the performance of a downstream task with less training data (e.g. object detection [23–25], segmentation [26, 27] or style transfer [28, 29]) and obtained state-of-the-art results. Recently, to reduce the dependency on large-scale labeled datasets, self-supervised pre-training has been gaining attention. Common choices of the pretext task include contrastive objectives based on data augmentation [30, 31] or masked autoencoding [32]. In natural language processing, learning rich representation through an *self-supervised* language model objective [33–35] has also been shown to boost performance.

In speech processing, early applications of pre-training deep neural networks mainly lay in automatic speech recognition (ASR), with the main goal of speeding up optimization and reducing generalization error [36, 37]. In recent years, unsupervised or self-supervised speech representation learning utilizing massive, unlabeled speech data has become a popular research topic [38, 39].

1.3 Thesis Overview

This thesis focuses on addressing the three problems described in Section 1.2 with the power of pre-training. Figure 1.5 shows the overall scope of this thesis. First, the fundamentals of VC are presented in Chapter 2. A more detailed review of the two current mainstream approaches for VC, namely seq2seq modeling for parallel VC and rec-syn-based non-parallel VC, will be presented. Common evaluation metrics and protocols will also be introduced.

To solve the first problem described in Section 1.2.1, in Chapter 3, we propose a pre-training technique for seq2seq VC modeling to reduce the requirement of the parallel

dataset size. Specifically, we identify that ASR and TTS, two of the most studied tasks in speech processing, have a common property of a comparatively pure linguistic hidden latent space which is a suitable knowledge source of transfer for VC. With a novel two-stage pre-training/fine-tuning framework, we show that it is possible to train a seq2seq VC model with similar quality using only 5 minutes of parallel data, which is a much-relaxed condition compared to the commonly adopted 1-hour setting.

To solve the second problem described in Section 1.2.2, in Chapter 4, we explore the possibility of applying self-supervised speech representations (S3Rs) to non-parallel VC. We contributed to the open-source activity by releasing a new toolkit for VC called S3PRL-VC, an extension of the self-supervised speech pre-training and representation learning (S3PRL) toolkit. We present a comparative study of S3R-based VC in various aspects, including the synthesizer model type, different VC tasks, supervision, and discretization. We also show a comparison with the state-of-the-art VC systems in VCC2020 and demonstrate the room for improvement.

To solve the third problem described in Subsection 1.2.3, we propose to combine both seq2seq modeling and non-parallel frame-based VC methods. In Chapter 5, we first present an idea to tackle VC problems where the training target is impossible to collect. The idea is to use a seq2seq model to first modify the temporal structure which is strongly related to the content-related attributes, then use a non-parallel VC method to restore the speaker identity. We verify the possibility of applying such an idea to dysarthric VC and showcase initial investigation results. In Chapter 6, we further explore two other methods that also combine seq2seq modeling and non-parallel frame-based VC, and evaluate in total three methods in a unified setting on the task of FAC. In our experiments, we found that no single method was superior to the others, and more importantly, we show the difficulties in the evaluation of atypical

VC applications. Finally, we conclude the thesis in Chapter 7.

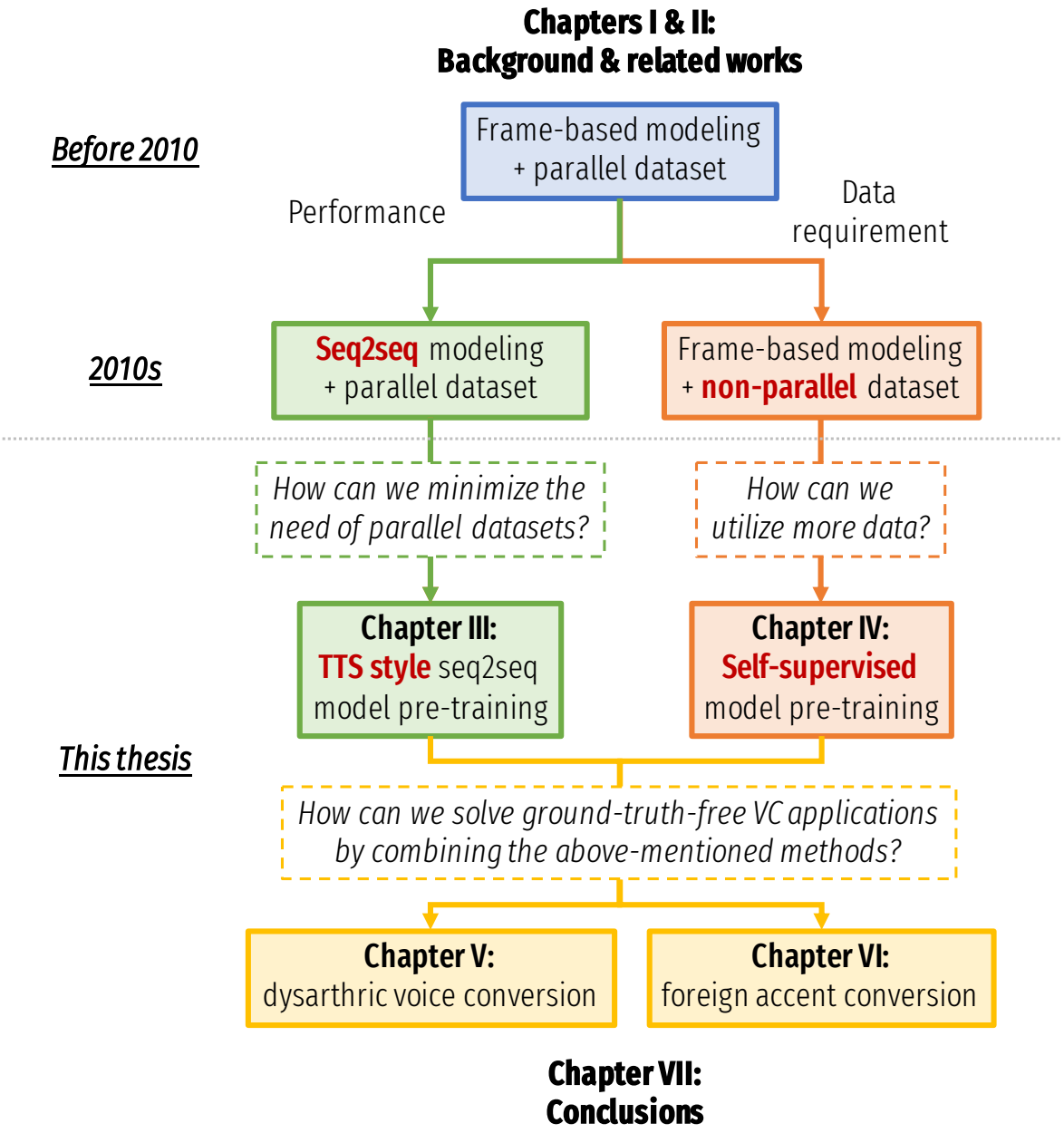


Figure 1.5: *Thesis scope.*

2 Background and related Work

In this chapter, we provide background materials to understand the main contents of this thesis. In Section 2.1, sequence-to-sequence modeling for voice conversion with parallel data, including the training objectives and common model architectures, is first introduced. In Section 2.2, two lines of representative VC approaches that utilize non-parallel datasets with a frame-based model are covered. Note that this review can be limited and not fully comprehensive, but it should be sufficient for readers to understand the remainder of this thesis. Finally, in Section 2.3, commonly used evaluation metrics and protocols for VC are described.

2.1 Voice conversion based on sequence-to-sequence modeling with parallel data

Generally speaking, seq2seq models aim to learn a mapping between a source feature sequence $\mathbf{X} = \boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and a target feature sequence $\mathbf{Y} = \boldsymbol{y}_{1:m} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m)$ which are often of different length, i.e, $n \neq m$ [40]. In speech processing, given a one-dimensional speech signal, it is a common practice to represent it as a sequence of acoustic feature vectors, where each acoustic feature vector is calculated from a speech frame given by a pre-defined frame size and frame shift. A commonly adopted choice is log mel-spectrogram. Thus, when applying seq2seq modeling to VC,

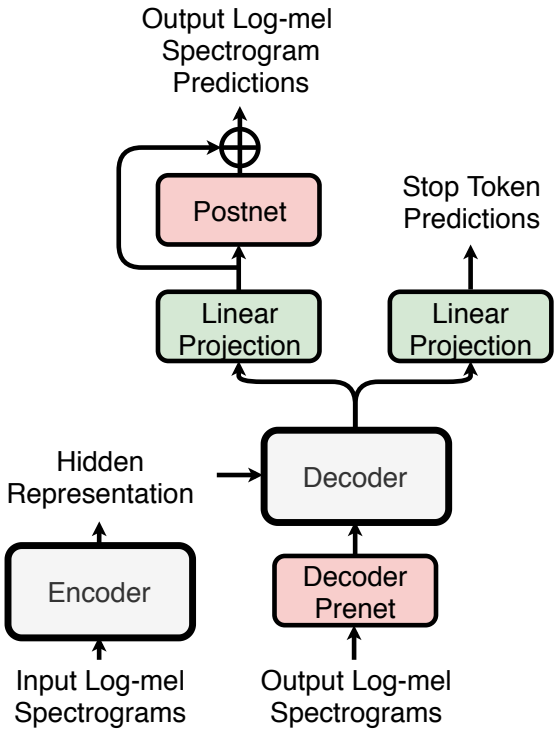


Figure 2.1: *Illustration of the architecture of a sequence-to-sequence VC model.*

the source and target feature sequences are exactly the source and target acoustic feature frames extracted from the respective speech utterance signals. In the remainder of this thesis, unless specified, the input and output of the seq2seq VC models refer to the acoustic feature frames instead of the original speech signals.

2.1.1 General architecture and training objectives

Figure 2.1 shows a general architecture of a seq2seq VC model, which is of an encoder—decoder structure [40]. The encoder (Enc) first maps the input feature sequence $\boldsymbol{x}_{1:n}$ into a sequence of hidden representations:

$$\mathbf{H} = \boldsymbol{h}_{1:n} = \mathrm{Enc}(\boldsymbol{x}_{1:n}). \tag{2.1}$$

The original seq2seq modeling framework [40] was *autoregressive* (AR), which means that during the output process, to generate the current output $\boldsymbol{y}_t$ at time step t, the decoder (Dec) takes, additional to the encoder output, i.e. the hidden representations $\boldsymbol{h}_{1:n}$, the previously generated features $\boldsymbol{y}_{1:t-1}$ (some also call them "historical information")into consideration:

$$\hat{\boldsymbol{y}}_t = \text{Dec}(\boldsymbol{h}_{1:n}, \hat{\boldsymbol{y}_{1:t-1}}). \tag{2.2}$$

When applying seq2seq models to tasks where the output is discrete, for example, machine translation or speech recognition, the output of the decoder at each time step is a $K+1$-way probability vector where K is the vocabulary size and the extra entry is for a special *stop token* that terminates the decoding process if generated. However, the same mechanism cannot be applied to the task of VC since the outputs are acoustic feature vectors of continuous values. Therefore, similar to many seq2seq text-to-speech (TTS) models [41,42], the decoder outputs pass through two separate linear projection layers, such that one outputs vectors of the same dimension as the acoustic features, and the other outputs a scalar value that represents the probability that the decoding process ends at the current time step.

The training objective contains two terms. The first term is an element-wise regression loss. Although many researchers have used an L1 loss, an L2 loss, or a combination of them, there has not been evidence on which choice is superior to the others. The second term is a binary cross-entropy loss for the stop token prediction. In addition, teacher-forcing is a common practice for training seq2seq VC models. That is, in Equation 2.2, the "previous generated vectors" during the decoding process are the corresponding ground truth target acoustic feature vectors. Such a practice can accelerate the training process by avoiding the actual AR decoding process.

Some extra components and techniques are adopted in the seq2seq model to improve

performance and stabilize training, most of which are inspired by the success of modern seq2seq TTS models [41, 42].

- A prenet is usually added to the decoder, which serves as an information bottleneck essential for training the AR process.
- To compensate for the missing future context information in the AR decoder, a five-layer CNN postnet is used to predict a residual that is added to the projected output.
- Introducing the reduction factor r greatly helps speed up convergence and reduce training time and memory footprint. Specifically, at each decoding step, r non-overlapping frames are predicted. Since adjacent speech frames are often correlated, this technique allows the decoder to correctly model the interaction with the hidden representation sequence.

2.1.2 Model architectures

In this subsection, we describe two commonly used model architectures in seq2seq VC modeling.

Recurrent neural network based model

The recurrent neural network (RNN) was adopted in the very first seq2seq modeling paper [40] as it is a straightforward choice for seq2seq modeling due to its ability to model long-range context. A representative RNN-based seq2seq VC model is ATTS2S-VC [16], which is based on the Tacotron2 TTS model [42]. Figure 2.2 shows the encoder and decoder structures.

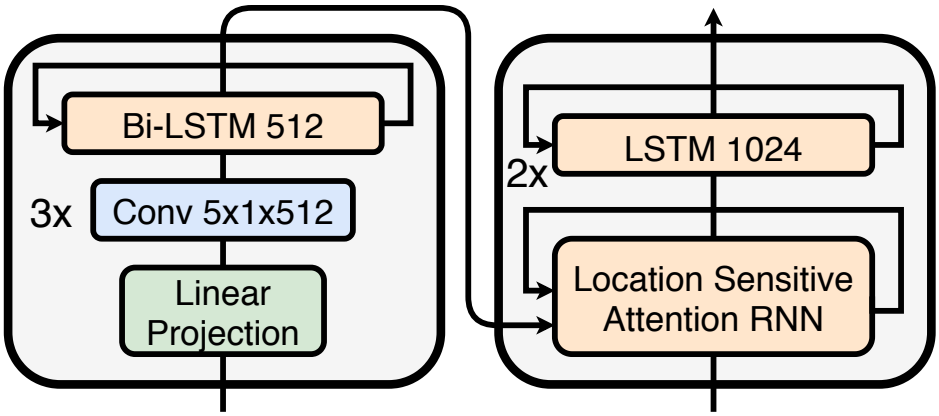


Figure 2.2: *RNN-based encoder and decoder in seq2seq VC.*

The encoder first linearly projects the input log-mel spectrogram, followed by a stack of convolutional layers, batch normalization, and ReLU activations. The output of the final convolutional layer is then passed into a bi-directional LSTM layer to generate the hidden representations.

For each decoder output step, an attention mechanism [43, 44] is used to attend to different positions of the hidden representation sequence. First, a context vector $\boldsymbol{c}_t$ is calculated as a weighted sum of $\boldsymbol{h}_{1:n}$, where the weight is represented using an attention probability vector $\boldsymbol{a}_t = (a_t^{(1)}, \cdots, a_t^{(n)})$. Each attention probability $a_t^{(k)}$ can be thought of as the importance of the hidden representation $\boldsymbol{h}_k$ at the current time step. As in Tacotron2, the location-sensitive attention is adopted [45], which takes cumulative attention weights from previous decoder time steps as an additional feature to encourage forward consistency to prevent repeated or missed phonemes. The context vector is then concatenated with the prenet output and passed into a stacked uni-directional LSTM network to predict the r output frames. The procedure mentioned

above can be formulated as follows:

$$\boldsymbol{a}_t = \text{attention}(\boldsymbol{q}_{t-1}, \boldsymbol{h}_{1:n}), \tag{2.3}$$

$$\boldsymbol{c}_t = \sum_{k=1}^{n} a_t^{(n)} \boldsymbol{h}_k, \tag{2.4}$$

$$\boldsymbol{y}_t, \boldsymbol{q}_t = \text{Dec}(\boldsymbol{y}_{1:t-1}, \boldsymbol{q}_{t-1}, \boldsymbol{c}_t). \tag{2.5}$$

In the following subsubsections, we describe two additional losses that are used in the ATTS2S-VC framework, with the motivation to further stabilize the training of the seq2seq model.

Guided Attention Loss The guided attention (GA) loss was first introduced in seq2seq TTS [46]. The motivation is that for seq2seq speech synthesis tasks like TTS and VC, the attention alignment is usually monotonic and linear. Therefore, encouraging the attention matrix to be diagonal can speed up attention learning and convergence.

The GA loss assumes that the i-th element in the input feature sequence progresses nearly linearly concerning the j-th element of the output feature sequence, i.e., $i \sim \alpha j$, where $\alpha \sim \frac{n}{m}$. Therefore, the attention matrix $\mathbf{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m]$ should be a nearly diagonal. One may therefore define a penalty matrix $\mathbf{G}$, where the i, j-th element $g_{i,j}$ is defined:

$$g_{i,j} = 1 - \exp\left\{\frac{-(\frac{i}{n} - \frac{j}{m})^2}{2\sigma_g^2}\right\}, \tag{2.6}$$

where σ_g controls how close $\mathbf{A}$ is to a diagonal matrix. The guided attention loss $\mathcal{L}_{\text{ga}}$ is then defined as

$$\mathcal{L}_{\text{ga}} = \lambda_{\text{ga}} ||\mathbf{G} \odot \mathbf{A}||_1, \tag{2.7}$$

where $\odot$ indicates an element-wise product and λ_{ga} is the weight for the guided attention loss.

Context Preservation Loss The context preservation (CP) loss was first proposed in [16] with the motivation to maintain linguistic consistency after conversion. Specifically, to encourage the source encoder to generate meaningful hidden representations, two additional networks are introduced as the context preservation mechanism: a source decoder SrcDec for reconstructing the source feature sequence from the hidden representations, and a target decoder TarDec for predicting the target feature sequence from the context vectors, $\mathbf{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_m]$:

$$\tilde{\mathbf{X}} = \text{SrcDec}(\mathbf{H}), \tag{2.8}$$

$$\tilde{\mathbf{Y}} = \text{TarDec}(\mathbf{C}). \tag{2.9}$$

The context preservation loss is then defined as:

$$\mathcal{L}_{\text{cp}} = \lambda_{\text{cp}}(||\tilde{\mathbf{X}} - \mathbf{X}||_1 + ||\tilde{\mathbf{Y}} - \mathbf{Y}||_1), \tag{2.10}$$

where λ_{cp} is the weight for the context preservation loss.

Transformer-based model

The Transformer [47] refers to a type of DNN architecture that relies solely on feed-forward network (FFN) blocks and attention blocks, and has gained much success in many speech processing tasks [48]. Here we describe a representative Transformer-based seq2seq VC model named the Voice Transformer Network (VTN) [49]. In the following subsubsections, we first describe the overall structure of VTN, then we describe core components used in a typical Transformer model.

Overall structure Figure 2.3 shows the structure of VTN. The encoder in a typical Transformer model is composed of a stack of encoder *blocks*, where each encoder block consists of a multi-head attention (MHA) sublayer and an FFN sublayer, followed by

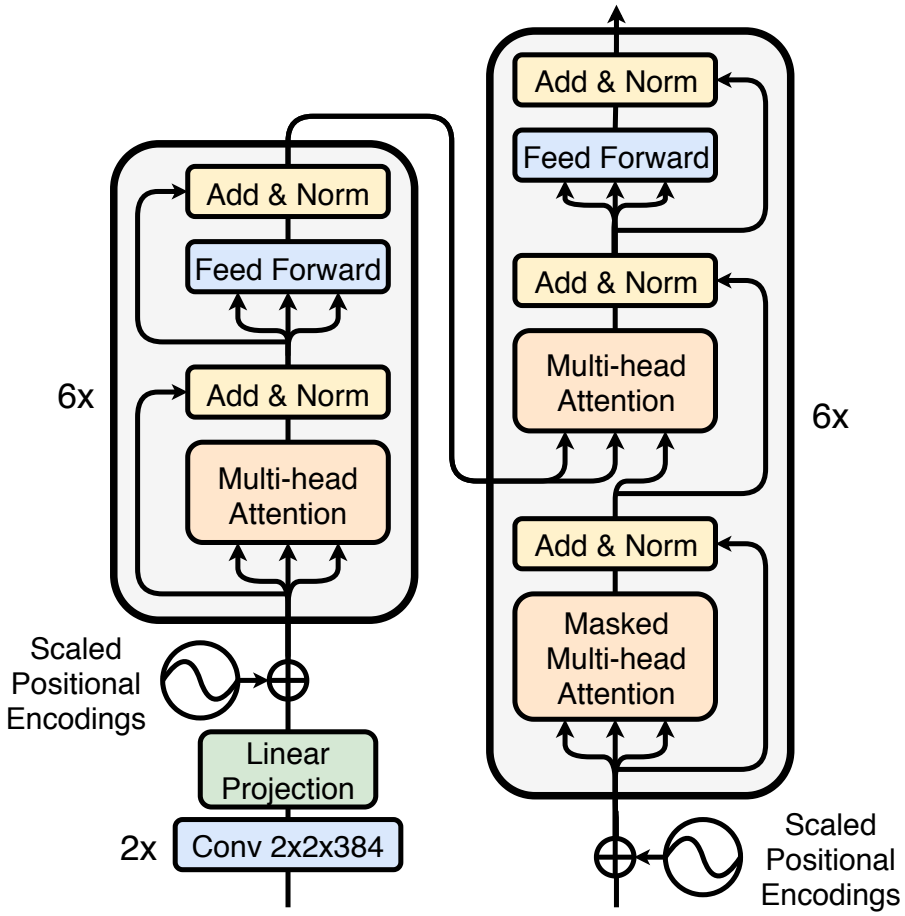


Figure 2.3: *Transformer-based encoder and decoder in seq2seq VC.*

a residual connection and layer normalization [50]. The MHA layers in the encoder are *self-attention* layers since the queries, keys, and values are all from the output from the previous layer.

The encoder in VTN is slightly different from that of the original Transformer since the latter was originally designed for the task of machine translation, whose input was discrete. For speech processing tasks where the input is speech (typically represented by a sequence of acoustic features), such as speech recognition or VC, it is a common practice to downsample the input. As in [51], a stack of two convolutional layers

with stride 2×2 is used to downsample the input by a fraction of 4. Doing so can not only reduce the memory footprint but also speed up attention convergence and approximate phoneme-level or even character-level linguistic contents [17]. The output of the convolutional stack is then sent through a linear projection layer. Then, a scaled positional encoding (SPE) is added, and then the output is sent to the encoder.

The decoder is also a stack of identical decoder blocks as in the encoder. In each decoder block, the first sublayer is a *masked* self-attention MHA sublayer, where a mask is utilized such that at time step t, only vectors with time index up to and including t can be accessed. This preserves the AR property of the model. Then, an MHA sublayer uses the outputs from the previous layer as queries and $\mathbf{H}$ as the keys and values, which ensembles the encoder—decoder attention described in Section 2.1.2. Finally, an FFN sublayer is used, as in the encoder. Again, all sublayers are wrapped with a residual connection and layer normalization.

Core components in the Transformer model Below we describe the core components used in Transformer models. In a Transformer model, a hyper-parameter d_{model} needs to be first decided, which in general defines the size of the model.

Multi-head attention (MHA) sublayer. An MHA layer is defined as:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\text{head}_1, \cdots, \text{head}_h]\mathbf{W}^O, \tag{2.11}$$

$$\text{head}_i = \text{Att}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V), \tag{2.12}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the input matrices that, following [47], are referred to as the query, key, and value, respectively. MHA uses h different, learned linear projections $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ to map the inputs to different *heads*, and then perform the Att operation in parallel. The outputs from all heads are concatenated and projected with $\mathbf{W}^O$. As

in [47], the Att operation is implemented scaled dot-product attention is used:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_{att}}})\mathbf{V}, \tag{2.13}$$

where d_{att} is the attention dimension.

Position-wise feed-forward network (FFN) layer. An FFN layer is defined as:

$$FFN(\boldsymbol{x}) = \max(0, \boldsymbol{x}\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2, \tag{2.14}$$

which is independently applied at each time step (position) with different parameters from layer to layer.

Layer normalization and residual connection. Around either of the above-mentioned sublayers, a residual connection followed by layer normalization [50] is employed. For input X of a sublayer, the output is given as:

$$\text{LayerNorm}(\mathbf{X} + \text{Sublayer}(\mathbf{X})). \tag{2.15}$$

Due to the residual connections, all sublayers have the same output dimension d_{model}.

Scaled positional encoding (SPE). In the original Transformer [47], since no recurrent relation is employed in the Transformer, to let the model be aware of information about the relative or absolute position of each element, the triangular (sinusoidal) positional encoding (PE) [52] is added to the inputs to the encoder and decoder. In this work, we adopt the SPE [53], which is a generalized version of the original PE that scales the encodings with a trainable weight α, so that they can adaptively fit the scales of the encoder and the decoder:

$$\text{SPE}(t) = \begin{cases} \alpha \cdot \sin(\frac{t}{10000^{\frac{2t}{d_{\text{model}}}}}), & \text{if } t \text{ is even,} \\ \alpha \cdot \cos(\frac{t}{10000^{\frac{2t}{d_{\text{model}}}}}), & \text{if } t \text{ is odd.} \end{cases} \tag{2.16}$$

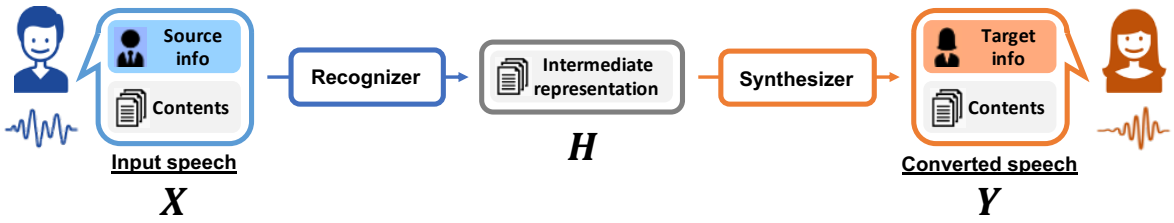


Figure 2.4: *A voice conversion framework based on the information perspective, which allows training with non-parallel data.*

2.2 Voice conversion based on frame-based modeling with non-parallel data

Collecting a parallel corpus for VC training is expensive in terms of time, money, and human labor. Therefore, researchers have been studying non-parallel VC extensively. A line of work makes use of a *reference speaker* to record training utterances that are parallel to those of the source and target speakers [54–59]. Another line of work is based on the recent success in cycle-consistency training [60, 61].

This section focuses on a particular line of work that reflects the basic definition of VC, which is to preserve the linguistic contents in the source speech. Figure 2.4 illustrates the conversion process. Starting from the source speech $\mathbf{X}$, a recognizer (or encoder) can first be used to extract the spoken contents, $\mathbf{H}$. The extracted output is referred to as the intermediate representation. It is then consumed by the synthesizer (or decoder) to generate the converted speech, $\mathbf{Y}$. The following equation describes such a process.

$$\mathbf{Y} = \text{Synth}(\mathbf{H}), \mathbf{H} = \text{Recog}(\mathbf{X}). \tag{2.17}$$

It is of most importance that the recognizer can extract the essential information that one wishes to preserve, depending on the actual application. For instance, in

Table 2.1: *A comparison of the autoencoding-based framework and the recognition-synthesis-based framework for voice conversion with non-parallel data.*

Framework	Optimization of the recognizer and the synthesizer	How to extract essential information
Autoencoding	Jointly	Proper information bottleneck
Recognition-synthesis	Separately and sequentially	Supervision or proper training objectives

speaker conversion, the intermediate representation should contain mostly linguistic content but not any source speaker information. Based on such a representation, the synthesizer is expected to inject the target speaker's identity and generate the converted speech. Now if the representation, unfortunately, contains remaining source speaker information, then the target speaker information injected by the synthesizer will be messed up, hurting the final conversion similarity. On the other hand, such a speaker-independent representation may not be desired in the application of foreign accent conversion, since the source speaker identity should be preserved.

The focus of this line of research then becomes **how to disentangle the spoken contents along with the desired attributes from the other factors in speech.** The efforts dedicated by the past researchers can be categorized into two types: the **autoencoding-based** approach, and the **recognition-synthesis (rec-syn)-based** approach[1]. Table 2.1 summarizes the two biggest differences. In Section 2.2.1, the autoencoding-based approach is first introduced. In Section 2.2.2, the rec-syn-based approach is introduced.

[1]The term *recognition-synthesis* was first defined in [62], which was only referred to VC systems composed by an ASR model and a speaker-dependent synthesizer. In this thesis, it is defined to be any VC system that separately trains the recognizer and synthesizer.

Figure 2.5: *Illustration of an autoencoding-based speaker voice conversion framework.*

2.2.1 Autoencoding-based approach

In machine learning, an autoencoder (AE) is a type of DNN that learns a useful, efficient coding of unlabeled data. An AE has the following properties:

- An AE is composed of an encoder that transforms the input data into an encoding, and a decoder that tries to reconstruct the input data from the encoded representation.

- A vanilla AE only uses a reconstruction objective to train the encoder and decoder jointly. As a result, the model might accidentally learn to just copy the input data, such that perfect reconstruction is achieved (i.e., the training loss becomes zero), and no useful feature is learned. To avoid this tendency, certain regularization techniques are necessary.

The above-mentioned properties exactly match the framework described at the beginning of Section 2.2. It is therefore a naive and straightforward idea to apply such a framework to VC.

AutoVC [63] is a representative AE-based VC model. Figure 2.5 shows an illustration of the structure of AutoVC when applied to speaker conversion. The roles of the encoder and decoder are essentially the same as the roles of the recognizer and synthesizer described previously, respectively. It is worthwhile mentioning that to

train an AE-based VC model for speaker conversion, it is a common practice to use a multi-speaker training corpus. If the hidden representation is ideal, i.e., contains no speaker information, then it would be impossible for the decoder to reconstruct the input speech data with such a feature alone. It is therefore necessary to supply the decoder with a *speaker representation*. Another interpretation is that, by providing the decoder such a clue about the speaker, the hidden representation is encouraged to be free from speaker information, in consideration of encoding efficiency. Another note is that, as a result of using a multi-speaker training corpus, the encoder is considered *speaker-independent*, which means that it can effectively encode speech from speakers unseen during training.

AutoVC proposed to *carefully* tune the dimensionality of the hidden representation. Others have proposed a variety of ways to regularize the encoder, including variational autoencoding [64–66], vector quantization [67–69], instance normalization [70], and hand-crafted speech related constraints [71].

A major drawback of autoencoding-based approaches is the trade-off between reconstruction accuracy and generalization ability. As mentioned earlier, the reconstruction loss could easily hit zero by learning an identical mapping, which is not desired since the encoder then does not disentangle any information, including those that should be discarded. On the other hand, sacrificing reconstruction ability leads to sub-optimal quality, resulting in blurry and unclear speech. As a result, in the voice conversion challenge (VCC) 2020 [9], the best-performing autoencoding-based system only ranked 13-th out of the 31 teams in task 1 and ranked 9-th out of the 28 teams in task 2.

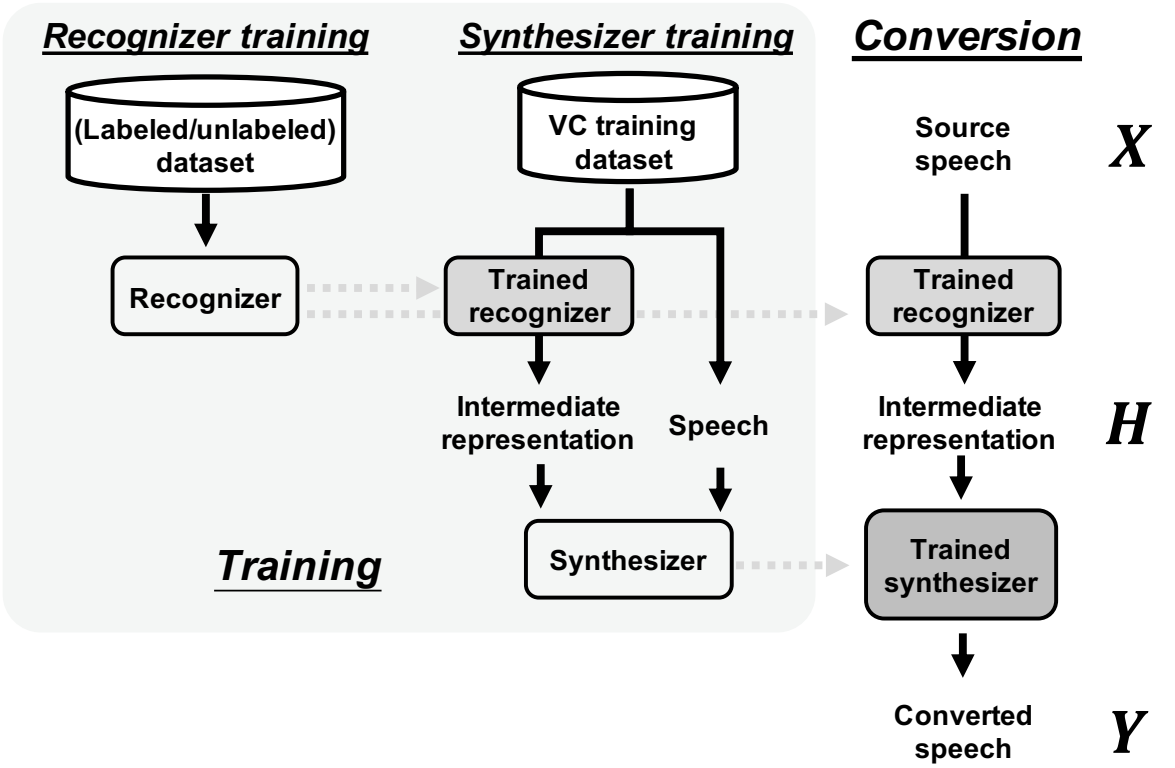


Figure 2.6: *The training and conversion procedures of recognition-synthesis based voice conversion.*

2.2.2 Recognition-synthesis-based approach

Rec-syn-based VC takes another approach and optimizes the recognizer and synthesizer not only separately but *sequentially*. Figure 2.6 illustrates the training and conversion processes. The recognizer is first trained on a pre-training dataset, which often contains multi-speakers, as in autoencoding-based VC. After the recognizer is trained, the synthesizer is then trained to reconstruct speech from the intermediate representation, which is extracted using the trained recognizer. In the conversion phase, the converted speech is generated following Eq. 2.17. The recognizer takes the source speech as input and extracts the intermediate representation, which is consumed

by the synthesizer to generate the converted speech.

Rec-syn-based VC has been gaining attention, mostly due to its superior performance compared to that of autoencoding-based VC. In VCC2020, all the systems that outperformed the best-performing autoencoding-based VC system were rec-syn-based. One hypothesis is that rec-syn-based VC benefits from the decoupling of the recognizer and synthesizer training: the training of the recognizer is to extract the desired representation, and the synthesizer is trained to optimize the reconstruction quality. Another advantage is that the training data of the recognizer and the synthesizer can be different. In autoencoding-based VC, because of the reconstruction loss, only high-quality datasets like the VC training set can be used. On the other hand, in rec-syn-based VC, while the synthesizer training is also limited to the high-quality VC training set, there is no such constraint on the recognizer, such that it can be trained on a massive dataset that is much larger than the VC training set. It is believed that the increased data and the sole objective function of the recognizer training form a stronger information bottleneck for preserving the linguistic contents, compared to those used in the autoencoding-based framework.

Efforts dedicated to this line of research can be divided into two directions. The first direction is to improve the reconstruction quality by adopting better generative modeling techniques, such as generative adversarial networks (GANs) [72], normalizing flows [73], variational inference [64], diffusion modeling [74, 75], or a combination of them.

The second direction develops different types of intermediate representations, which results in different recognizer designs and training schemes. In the literature, many types of intermediate representations have been used, all of which have their respective pros and cons. Table 2.2 presents a comparison of the features based on various aspects.

Table 2.2: *A comparison of intermediate representations in recognition-synthesis-based voice conversion.*

Representation	Text	Phonetic Posteriorgram	Self-supervised speech representations
Extractor	ASR model		self-supervised model
Training data	labeled data		unlabeled data
Resolution	token level	frame level	
Continuous?	discrete	continuous	can be either
Examples	[76, 77]	[78–80]	[81–86]

In the following, we introduce three widely used categories.

Text

Text is a straightforward choice, as one can simply concatenate a pre-trained ASR and text-to-speech (TTS) model. In VCC2020, one of the baseline systems called ASR+TTS [76] and the top system of the intra-lingual task [77] both adopted text as the intermediate representation and achieved outstanding performance in terms of similarity. This is mainly owing to the discrete and token-level nature of text. Since prosodic information including the speaking rate and the pitch pattern are discarded after recognition, the synthesizer needs to use a powerful model like a seq2seq network to reconstruct the target characteristics. However, this approach suffers from mispronunciation when the accuracy of the ASR and TTS model is insufficient, as shown in [76]. There are also VC scenarios where the source style needs to be preserved, such as singing VC [87].

Phonetic posteriorgrams or bottleneck features

Phonetic posteriorgrams (PPGs) were first applied to VC in [88] represent the frame-wise posterior probabilities of each phonetic class, which are derived from the acoustic model (AM) of an HMM-based ASR model. The training target of the AM is phoneme labels, so only the output of the last layer of the AM has the physical meaning of PPG, but some have proposed to use the output from other layers. For example, the system in [17] used the output before the softmax layer and referred to them as bottleneck features (BNFs). Either PPGs or BNFs are frame-level continuous features, thus better preserving the linguistic contents and can help produce high-quality speech.

VC systems based on PPG or BNF have been showing their supremacy in recent VCCs. They first showed their power by raking first in VCC2018 [89]. Later on, in VCC2020, several top-performing systems also used such a feature [78–80]. It was also shown in the challenge results that many systems based on such a feature were top-ranking systems.

However, the frame-level nature makes the conversion of the speaking rate difficult. Efforts needed for the frame-level labels of the ASR dataset also raised the difficulty of constructing the system.

Self-supervised speech representations

To reduce the labeling cost of training ASR models, applying self-supervised learning (SSL) to VC has become increasingly popular. Being free from labeled data not only reduces the labeling cost but also makes it possible to use more unlabeled datasets and work under low-resource settings. SSL has been applied to a wide variety of VC settings, including any-to-one VC [82], many-to-many VC [81], any-to-any VC [83, 84] and cross-lingual VC [85].

The focus of Chapter 4 will be on SSL-based VC. A review of SSL-based speech representation learning as well as an in-depth study on SSL-based VC will be presented.

2.3 Common evaluation metrics and protocols in Voice conversion

In this section, common evaluation metrics and protocols in VC are introduced. For speech synthesis tasks such as TTS and VC, the gold standard for evaluation is listening tests, where human raters listen to samples generated by different methods and give their opinions. The reason why *subjective* test results are favored compared to *objective* metrics is that current objective metrics do not align well with human perception [90]. However, it is still helpful to use objective metrics to monitor the performance during model development, to reduce the cost of conducting subjective tests. In the following subsections, the subjective evaluation protocol is first described. Then, several objective evaluation metrics used in this work will be introduced.

Subjective evaluation methods

Two common dimensions are measured in a subjective test for evaluating a VC system.

- **Naturalness.** Naturalness refers to the degree of how natural the generated voice sample sounds. Usually, the upper bound is defined by the samples of the target speaker. In VC research, the term "naturalness" is often used instead of "quality" to distinguish it from other performance factors. The most commonly used protocol for naturalness in VC research [7–9] is the mean opinion

score (MOS) test [91]. Subjects were asked to evaluate the naturalness of the converted and natural speech samples on a scale from 1 (completely unnatural) to 5 (completely natural).

For example, in the voice conversion challenge (VCC) 2020, the following instruction is given:

> Listen to the following audio and rate it for quality. Some of the audio samples you will hear are of high quality, but some of them may sound artificial due to deterioration caused by computer processing. Please evaluate the voice quality on a scale of 1 to 5 from "Excellent" to "Bad." Quality does not mean that the pronunciation is good or bad. If the pronunciation of the English is unnatural but the sound quality is very good, please choose "Excellent."

They were then asked to rate how natural the speech sounded on a five-point scale: (1) Bad, (2) Poor, (3) Fair, (4) Good, and (5) Excellent.

- **Conversion similarity.** Conversion similarity refers to how similar the speaker identity of the generated voice sample is compared to the target speaker. A commonly used protocol is the VCC style test, which was adopted in the past VCCs [7–9]. Specifically, listeners are given a pair of speech utterances consisting of a speech sample from the target speaker and a converted speech sample. Then, they were asked to determine whether the pair of utterances could be produced by the same speaker, with 4-level confidence in their decision, i.e., sure or not sure.

  In the voice conversion challenge (VCC) 2020, the following instruction is given:

  > Please listen to the following two audio samples and rate them for

> speaker similarity. Please consider who is speaking according to the characteristics of the sound and then make a choice using a 4-level scale that varies from "Same (sure)" to "Different (sure)" to rate the speaker similarity of the two audio samples. Please do not consider the content or language to which you are listening.

They were then asked to rate the speaker similarity of the two samples on a four-point scale: (4) same speaker, absolutely sure, (3) same speaker, not sure, (2) different speaker, not sure, (1) different speaker, absolutely sure.

The results can be presented in two ways. The first way, which is used in VCCs, is to present the raw results in the form of stacked bar charts. The other approach is to show the combined percentage of (4) same speaker, absolutely sure, and (3) same speaker, not sure.

Objective evaluation metrics

In this thesis, the following objective evaluation metrics are used.

- **Mel cepstrum distortion (MCD) [92].** The MCD is a commonly used measure of spectral distortion in VC, which is based on mel-cepstral coefficients (MCCs). It is defined as:

$$\text{MCD}[dB] = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{K}(mcc_d^{(c)} - mcc_d^{(t)})^2}, \tag{2.18}$$

where K is the dimension of the MCCs and $mcc_d^{(c)}$ and $mcc_d^{(t)}$ represent the d-th dimensional coefficient of the converted MCCs and the target MCCs, respectively. In practice, MCD is calculated in an utterance-wise manner. A dynamic time warping (DTW) based alignment is performed to find the corresponding frame

pairs between the non-silent converted and target MCC sequences beforehand. In this thesis, the WORLD vocoder [93] for MCC extraction and silence frame decisions and set $K = 24$.

- **F0 root mean square error (F0RMSE).** F0RMSE refers to the RMSE between the F0 of converted speech and that of the reference target speech. Similar to the calculation of MCD, DTW-based is performed and we take only the non-silent frames into account.

- **Character/word error rate (CER/WER).** The CER/WER is a rough estimate of the intelligibility of the converted speech.

- **ASV**: This metric, along with the protocol for calculating such a metric, is used in some recent VC literature [83, 84]. It calculates the accept rate from a pretrained automatic speech verification model which measures whether the speaker identity is converted by calculating the cosine similarity using speaker embeddings [94]. Specifically, the cosine similarity of the d-vectors [95] extracted from each converted utterance and the corresponding reference is calculated. Then, the percentage of the testing utterances whose cosine similarity exceeds a pre-calculated threshold is reported.

2.4 Summary

In this chapter, fundamental background knowledge for understanding this thesis was reviewed. In Section 2.1, seq2seq VC modeling with parallel data was introduced, including two commonly used model architectures and the training objectives. In Section 2.2, non-parallel frame-based VC methods were introduced. Specifically,

two representative lines of approaches, namely autoencoding-based and rec-syn-based methods, were introduced. Finally, in Section 2.3, the commonly used objective evaluation metrics and subjective evaluation protocols for VC are described. The contents in this chapter will be frequently referenced in later chapters in this thesis, and it is expected that readers will develop a deeper understanding while reading the subsequent chapters.

3 Pre-training for Sequence-to-sequence Voice Conversion

In this chapter, a novel pre-training method for sequence-to-sequence seq2seq modeling for voice conversion (VC) is described. The goal is to reduce the parallel dataset requirement of seq2seq VC models. Specifically, a two-stage pre-training approach is proposed to transfer knowledge from text-to-speech (TTS) and automatic speech recognition (ASR). The benefits of transferring knowledge from these two tasks are two-fold: (1) the task nature of TTS and ASR makes both of them a proper source to transfer knowledge from, and (2) the abundant resource contributed by the research community makes it easily accessible. Experimental results show that the proposed pre-training method can (1) improve performance compared to training from scratch in a high-resource setting, and (2) greatly reduce performance degradation when the training data size is limited.

3.1 Introduction

Compared to traditional machine learning models used in VC such as Gaussian mixture models [96], DNN-based models contain more model parameters, and thus typically require a larger amount of dataset to train. As described in Section 1.2.1,

most seq2seq VC papers [15–18] report results by using roughly one hour of parallel data. However, it is commonly assumed that only several minutes of parallel data is accessible in VC [7–9]. As we will show in later sections, directly training seq2seq models with, for instance, five minutes of parallel data suffers from severe performance degradation. It is therefore desirable to develop a solution to reduce the parallel dataset requirement of seq2seq VC models.

As described in Section 1.2.4, there has been a significant amount of effort dedicated to investigating different pre-training methods and objectives for different speech processing tasks. Nonetheless, different pre-training objectives lead to different representations, and an effective objective for seq2seq VC is still unclear.

In this chapter, a pre-training technique to transfer knowledge from two speech processing tasks, namely TTS and ASR, is proposed. They are referred to as *TTS-oriented pre-training* and *ASR-oriented pre-training*, respectively. There are two reasons behind the use of these two tasks.

- **Resource.** ASR and TTS are among the speech processing fields that have drawn the most attention, and an abundant amount of resources have been made publicly available. One can even say that the great success enjoyed by ASR and TTS largely owes to the vast large-scale corpora contributed by the community. These rich resources are believed to benefit VC which suffers from data scarcity.

- **Task nature.** In recent years, ASR and TTS systems based on neural seq2seq models have been shown to outperform traditional methods [41,97]. It is believed that lying at the core of these models is the ability to generate effective intermediate representations, which facilitates correct attention learning that bridges the encoder and the decoder.

  Figure 3.1 shows a unified comparison of TTS, ASR, and VC from an information

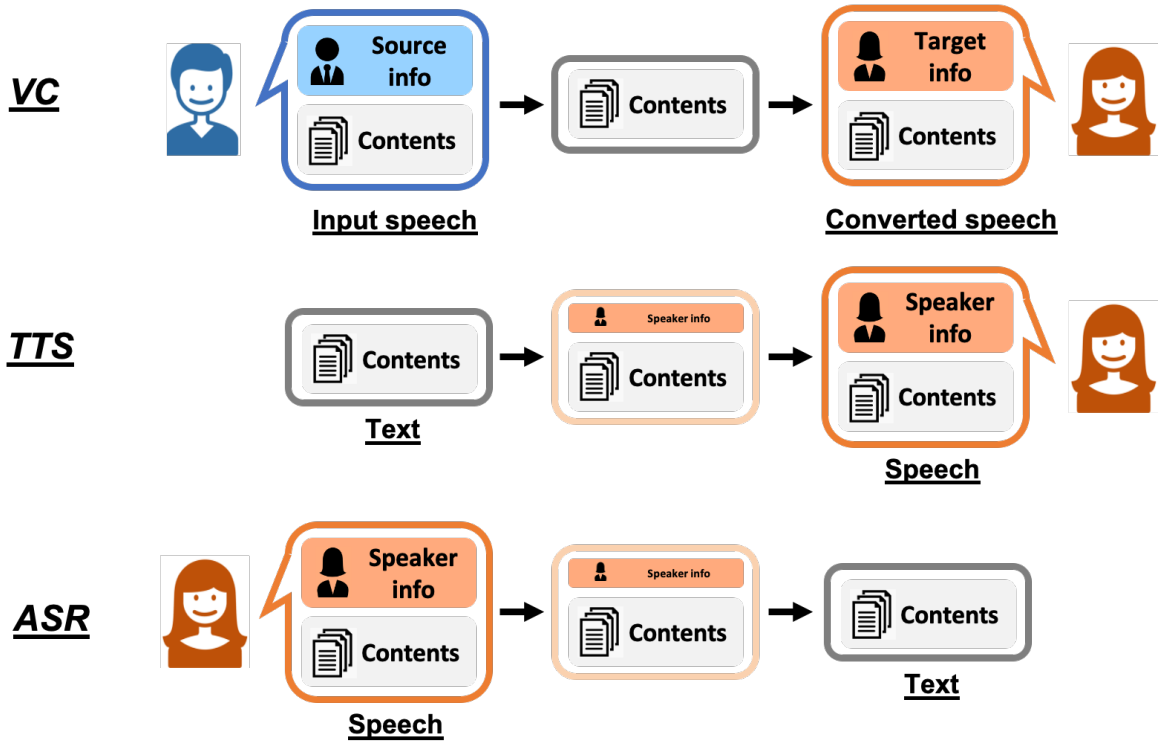


Figure 3.1: *Illustration of the relationship of VC, TTS, and ASR from an information perspective.*

perspective. Roughly speaking, speech consists of the linguistic contents and the speaker identity. The goal of VC is to remove the source speaker information from the source speech, and then inject the identity of the target speaker. Thus, a speaker-free intermediate feature space would be essential for a successful VC model, which is hard to facilitate given only a parallel corpus. On the other hand, TTS and ASR both aim to find a mapping between text and speech, as the former tries to add speaker information to the source while the latter tries to remove it. It is therefore suspected that the intermediate hidden representation spaces of these two tasks contain somewhat little speaker information and serve

as a suitable fit for VC.

The proposed method enjoys several advantages. First, it relies on *supervised* pre-training with well-defined speech processing objectives. As popular speech processing tasks (i.e. TTS and ASR) are adopted, large-scale datasets can be assumed easily accessible thanks to the vastly growing community. Also, the proposed method is flexible in that it needs neither the text label of the VC data nor carefully designed regularization methods, yet can still achieve great data efficiency. Finally, it is expected that the performance of pre-training would benefit from the rapid development of state-of-the-art models, thus improving the quality of the downstream VC task.

The contributions of this chapter are as follows:

- The TTS-oriented and ASR-oriented pre-training are proposed for seq2seq VC. Through systematical objective and subjective evaluations, it is shown that both are effective with sufficient data, while only TTS pre-training remains robust against the reduction of data.

- The hidden representation spaces of the learned models using different pre-training tasks are visualized, and their relationship to the performance is also shown.

- Two different model architectures for seq2seq VC, namely recurrent neural networks (RNNs) and Transformers, are examined. It is shown that the latter is superior to the former, which is consistent with the finding in most speech processing tasks [48].

3.2 Method

In seq2seq models for speech applications, effective intermediate representations can facilitate correct attention learning that bridges the encoder and the decoder, thus crucial to success. By the definition of VC, it is natural to try to encode the linguistic contents of the source speech into the hidden representations so that they can be maintained. Thus, it is conjectured that the core ability of successful seq2seq VC models is to generate and utilize high-fidelity hidden representations.

In theory, both TTS and ASR tasks aim to find a mapping between two modalities: speech and text. As speech signals contain all essential linguistic information, the hidden representation spaces induced by these two tasks should lie in the middle of the spectrum between speech and text. Thus, it is hypothesized such space is desirable for seq2seq VC models, and thus suitable for pre-training.

Figure 3.2 shows the proposed two-stage training procedure. In the first pre-training stage, a large-scale corpus is used to learn the initial seq2seq model parameters as a prior; then, in the second stage, the seq2seq VC model is initialized with the pre-trained model parameters and trained with a relatively smaller VC dataset. The goal of this pre-training procedure is to provide fast, sample-efficient VC model learning, thus reducing the data size requirement and training time. In addition, this setup is highly flexible in that (1) the speakers of the datasets used in pre-training and fine-tuning need not be the same, and (2) the utterances between the pre-training corpus and the VC dataset need not be parallel.

Let the parallel VC dataset be $\mathbf{D}_{\mathrm{VC}} = \{\boldsymbol{S}_{\mathrm{src}}, \boldsymbol{S}_{\mathrm{trg}}\}$, where $\boldsymbol{S}_{\mathrm{src}}, \boldsymbol{S}_{\mathrm{trg}}$ denote the source, target speech, respectively. Our goal is to find a set of prior model parameters to train the final encoder $\mathrm{Enc}^{\mathrm{S}}_{\mathrm{VC}}$ and decoder $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{VC}}$. In the following subsections, the details of the pre-training and fine-tuning processes are described.

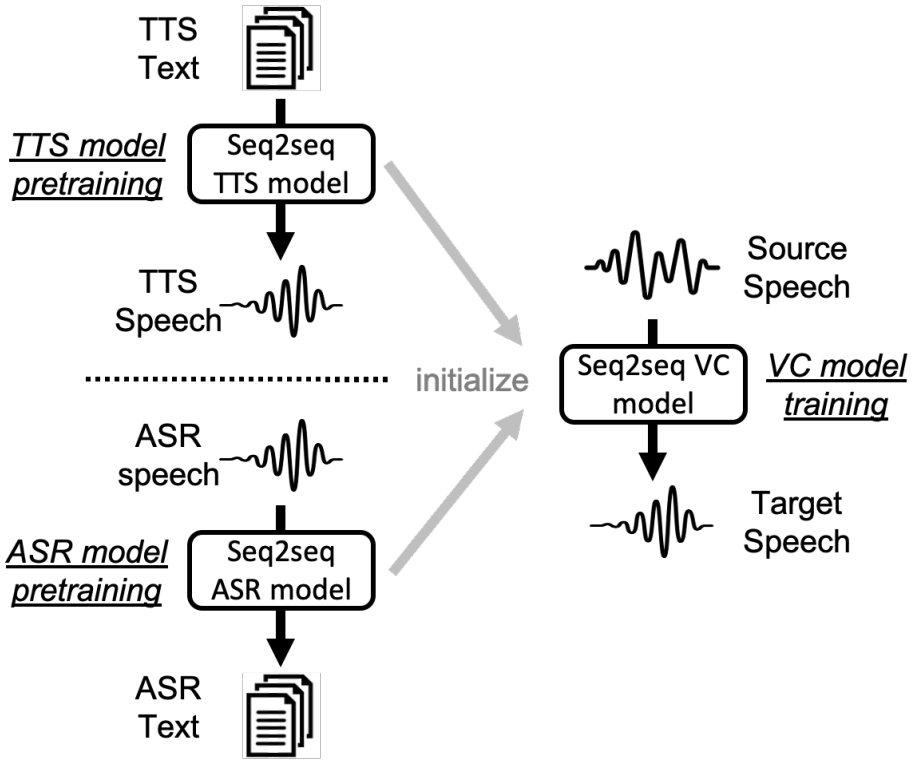


Figure 3.2: *Illustration of the concept of pre-training from seq2seq TTS or ASR to seq2seq VC.*

3.2.1 TTS-oriented pre-training

In TTS-oriented pre-training, it is assumed that access to a large *single-speaker* TTS corpus $\boldsymbol{D}_{\mathrm{TTS}} = \{\boldsymbol{T}_{\mathrm{TTS}}, \boldsymbol{S}_{\mathrm{TTS}}\}$ is available, where $\boldsymbol{T}_{\mathrm{TTS}}$, $\boldsymbol{S}_{\mathrm{TTS}}$ denote the text and speech of the TTS speaker respectively. The pre-training can be broken down into two steps.

A.1 *Decoder pre-training*: As in A.1 in Figure 3.3, the decoder is pre-trained, on $\boldsymbol{D}_{\mathrm{TTS}}$, by training a conventional TTS model composed of a text encoder $\mathrm{Enc}^{\mathrm{T}}_{\mathrm{TTS}}$ and a speech decoder $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{TTS}}$.

A.2 *Encoder pre-training*: Then, as in A.2 in Figure 3.3, the encoder is pre-trained,

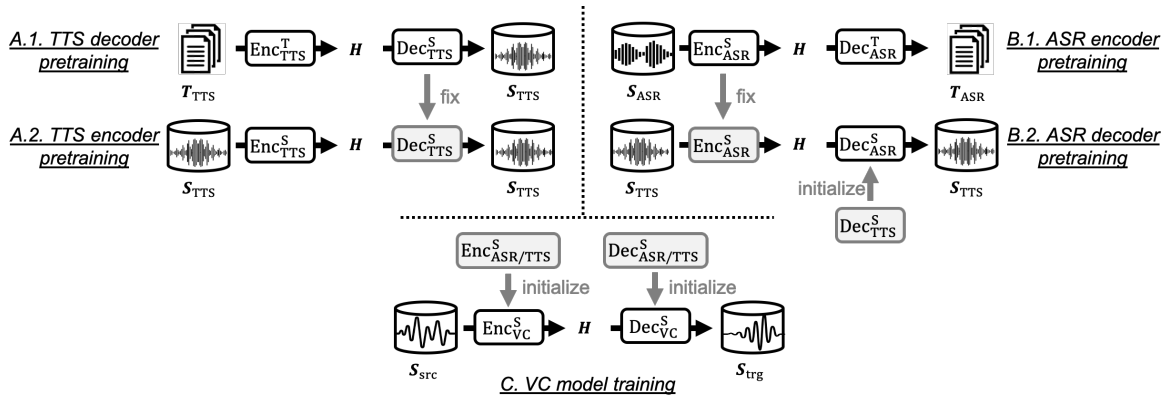


Figure 3.3: *Diagram of the pre-training procedures for TTS and ASR. Top left: TTS-oriented pre-training. Top right: ASR-oriented pre-training. Bottom: VC model training.*

also on the same $\boldsymbol{D}_{\mathrm{TTS}}$, by training an autoencoder which takes $\boldsymbol{S}_{\mathrm{TTS}}$ as input and output. The decoder here is the pre-trained $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{TTS}}$ and the parameters are fixed so that they are not updated during training. The desired pre-trained encoder $\mathrm{Enc}^{\mathrm{S}}_{\mathrm{TTS}}$ can then be obtained by minimizing the reconstruction loss.

The intuition of the encoder pre-training step is to obtain an encoder capable of encoding acoustic features into hidden representations that are recognizable by the pre-trained decoder. Another interpretation is that the final pre-trained encoder $\mathrm{Enc}^{\mathrm{S}}_{\mathrm{TTS}}$ tries to mimic the text encoder $\mathrm{Enc}^{\mathrm{T}}_{\mathrm{TTS}}$. In the first decoder pre-training step, since text itself contains pure linguistic information, the text encoder $\mathrm{Enc}^{\mathrm{T}}_{\mathrm{TTS}}$ is ensured to learn to encode an effective hidden representation that can be consumed by the decoder $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{TTS}}$. Fixing the decoder in the encoder pre-training process, as a consequence, guarantees the encoder to behave similarly to the text encoder, which is to extract fine-grained, linguistic-information-rich representations.

3.2.2 ASR-oriented pre-training

In ASR-oriented pre-training, it is assumed that a large *multi-speaker* ASR corpus $\boldsymbol{D}_{\text{ASR}} = \{\boldsymbol{S}_{\text{ASR}}, \boldsymbol{T}_{\text{ASR}}\}$ is available, where $\boldsymbol{S}_{\text{ASR}}$, $\boldsymbol{T}_{\text{ASR}}$ denote the speech and text data in $\boldsymbol{D}_{\text{ASR}}$, respectively. Similar to TTS-oriented pre-training, the ASR-oriented pre-training is again broken down into two steps.

B.1 *Encoder pre-training*: As shown in B.1 in Figure 3.3, the *encoder* is first pre-trained on $\boldsymbol{D}_{\text{ASR}}$ by training a conventional ASR model consisting a speech encoder $\text{Enc}^{\text{S}}_{\text{ASR}}$ and a text decoder $\text{Dec}^{\text{T}}_{\text{ASR}}$.

B.2 *Decoder pre-training*: The decoder pre-training procedure is depicted in B.2 in Figure 3.3. Note that the decoder pre-training here is performed on $\boldsymbol{D}_{\text{TTS}}$, rather than on $\boldsymbol{D}_{\text{ASR}}$. This is because $\boldsymbol{D}_{\text{ASR}}$ is a multi-speaker corpus, but the VC model architecture in this chapter focuses on one-to-one VC, i.e. modeling the conversion between one source speaker and one target speaker, thus cannot model individual speaker characteristics. Again, the decoder pre-training uses $\boldsymbol{S}_{\text{TTS}}$ as input and output, and the encoder is the pre-trained $\text{Enc}^{\text{S}}_{\text{ASR}}$ and kept fixed during training. To speed up convergence, the decoder is initialized with the one obtained in TTS decoder pre-training, namely $\text{Dec}^{\text{S}}_{\text{TTS}}$. The desired pre-trained decoder $\text{Enc}^{\text{S}}_{\text{ASR}}$ can then be obtained by minimizing the reconstruction loss.

The intuition of the ASR decoder pre-training differs from that of the TTS encoder pre-training. The ASR speech encoder $\text{Enc}^{\text{S}}_{\text{ASR}}$, trained with the ASR objective, should generate a compact hidden representation for decoding underlying linguistic contents. Such representations are believed to be easier to map to speech, thus suitable for pre-training the speech decoder $\text{Dec}^{\text{S}}_{\text{ASR}}$.

3.2.3 VC model fine-tuning

Finally, $\mathbf{D}_{\mathrm{VC}}$ is used to train the desired VC models $\mathrm{Enc}^{\mathrm{S}}_{\mathrm{VC}}$ and $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{VC}}$, with the encoder initialized with either $\mathrm{Enc}^{\mathrm{S}}_{\mathrm{TTS}}$ or $\mathrm{Enc}^{\mathrm{S}}_{\mathrm{ASR}}$, and the decoder with $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{TTS}}$ or $\mathrm{Dec}^{\mathrm{S}}_{\mathrm{ASR}}$, respectively. As we will show later, the pre-trained model parameters serve as a very good prior for adapting to the relatively scarce VC data, achieving significantly better conversion performance.

3.3 Experimental settings

3.3.1 Data

For $\mathbf{D}_{\mathrm{VC}}$, the CMU ARCTIC database [98] was used in the experiments, which contained parallel recordings of professional US English speakers sampled at 16 kHz. Data from four speakers were used: a male source speaker (*bdl*) and a female source speaker (*clb*), as well as a male target speaker (*rms*) and a female target speaker (*slt*). 100 utterances were selected for each validation and evaluation set, and the remaining 932 utterances were used as training data.

For $\boldsymbol{D}_{\mathrm{TTS}}$, a US female English speaker (*judy bieber*) from the M-AILABS speech dataset was used [99]. The sampling rate was 16 kHz. The training set contained 15,200 utterances, which was roughly 32 hours long.

For $\boldsymbol{D}_{\mathrm{ASR}}$, the LibriSpeech dataset [100] was used. The sampling rate was 16 kHz. The *train-clean-100* and *train-clean-360* sets were combined to get 460 hours of data from roughly 1170 speakers.

3.3.2 Implementation

The seq2seq model backbone was either the Voice Transformer Network (VTN) described in Section 2.1.2 or the RNN-based model described in Section 2.1.2. The experiment was carried out on the open-source ESPnet toolkit [101, 102], including feature extraction, training, and benchmarking. The official implementation has been made publicly available[1]. Interested readers are recommended to directly refer to the settings and configurations online. For the acoustic features, 80-dimensional mel filterbanks with 1024 FFT points and a 256-point frame shift were used as the acoustic features. The LAMB optimizer [103] was used, and the learning rate was set to 0.001.

3.3.3 Waveform synthesis module

To synthesize the final converted waveform from the generated acoustic features, the Parallel WaveGAN (PWG) [104] was used, which supported parallel, faster than real-time waveform generation[2]. For each target speaker, a speaker-dependent PWG was trained by conditioning on natural mel spectrograms extracted from the training data of each target speaker. Note that in the experiments, even for cases where only a subset of $\mathbf{D}_{\mathrm{VC}}$ can be accessed, the PWG trained with the full training dataset was still used, instead of training separate PWGs w.r.t. different training data sizes. This was because the goal of the experiment was to demonstrate the effects of various methods, instead of reflecting true scenarios.

[1]Originally the implementation was released as a part of ESPNET: `https://github.com/espnet/espnet/tree/master/egs/arctic/vc1`. Later on, an isolated toolkit was released at `https://github.com/unilight/seq2seq-vc`. Note that some settings of the latter, including the pre-training dataset and evaluation model, are slightly different from that in the former.

[2]The open-source implementation at `https://github.com/kan-bayashi/ParallelWaveGAN` was used.

3.3.4 Evaluation metrics

Subjective evaluation tests were conducted by following the protocols described in Section 2.3. They were performed using the open-source toolkit [105] which implemented the ITU-T Recommendation P.808 [106] for subjective speech quality assessment in the crowd using the Amazon Mechanical Turk (Mturk), and screened the obtained data for unreliable ratings. More than fifty listeners were recruited. A demo web page with samples used for subjective evaluation is available[3].

For objective evaluation metrics, the following were used: MCD, F0RMSE, and CER/WER. For definitions, please refer to Section 2.3. Here the F0RMSE is reported because seq2seq modeling can greatly improve prosody conversion. The ASR model used for calculating CER/WER was based on the Transformer architecture [51] and was trained using the LibriSpeech dataset [100]. The CER and WER for the ground-truth validation set were 0.9% and 3.8%, respectively, which could be regarded as the upper bound. Note that to avoid overfitting, the validation set MCD was used as the criterion for model selection, and the best-performing models proceeded to generate the samples for the subjective test.

3.4 Experimental evaluation results

3.4.1 Effectiveness of TTS-oriented pre-training on RNN and Transformer-based models

First, the effectiveness of TTS-oriented pre-training on not only Transformer-based (VTN) but also RNN-based seq2seq VC models was demonstrated. From the objective

[3] `https://unilight.github.io/Publication-Demos/publications/vtn-taslp/index.html`

evaluation results shown in Table 3.1, the following observations can be made.

- First, without pre-training, both VTN and RNN could not stay robust against the reduction of training data. The performance dropped dramatically with the reduction of training data, where a similar trend was also reported in [62]. This result identified the data efficiency problem of seq2seq VC.

- By incorporating TTS-oriented pre-training, both VTN and RNN exhibited a significant improvement in all objective measures, where the effectiveness was robust against the reduction in the size of training data. With only 80 utterances, both models could achieve comparable performance to that of using 932 training utterances except the F0RMSE, where in the case of VTN, the intelligibility was even better.

From the subjective evaluation results in Table 3.2, the following observations were obtained.

- Without pre-training, the VTN and RNN suffered from about 1.2 and 0.8 MOS points drop when the training data reduced from 932 to 80 utterances.

- With TTS-oriented pre-training applied, the naturalness of VTN and RNN improved by more than 1 point with 932 utterances and more than 2 points with 80 utterances.

- When the training data was reduced, there was only a very limited performance drop. These results demonstrated the effectiveness of the TTS-oriented pre-training technique.

3.4.2 Comparison of TTS-oriented and ASR-oriented pre-training

Next, the effectiveness of TTS-oriented and ASR-oriented pre-training was compared. The following observations could be obtained from Tables 3.1 and 3.2.

- With the full training set, ASR-oriented pre-training could bring almost the same amount of improvement compared to TTS-oriented pre-training.
- As the size of the training data reduces, the performance of the ASR-oriented pre-trained model dropped significantly, except for F0RMSE. This showed that ASR-oriented pre-training lacked the robustness essential for practical VC.

To investigate the failure of ASR-oriented pre-trained models against limited training data, the ASR result of a randomly-picked converted sample from the evaluation set using TTS-oriented and ASR-oriented pre-trained VTNs is shown in Table 3.3. Although TTS-oriented pre-training could not ensure complete linguistic consistency, the errors were minor and possibly due to the imperfect ASR engine used for evaluation, thus the result seemed reasonable. On the other hand, the recognition result of the ASR-oriented pre-trained model with 80 utterances had no connection to the source sentence. It is concluded that linguistic consistency was poorly maintained under the limited data scenario using ASR-oriented pre-training.

3.4.3 Comparison of RNN and Transformer based models

As an ablation study, the performance of RNN-based and Transformer-based models was compared by applying TTS-oriented pre-training to both models. Tables 3.1 and 3.2 gave the following observations.

- Without TTS-oriented pre-training, VTNs were less robust to training data reduction than RNNs in terms of objective measures but better in terms of sub-

jective measures. This was possibly because a more complex model like VTN is capable of generating better-sounding voices while being more prone to overfitting since it lacks attention regularizations such as the location-sensitive location, as suggested in [101].

- With TTS-oriented pre-training, it could be observed that VTNs outperformed RNNs in terms of all objective measures except F0RMSE and subjective scores. This was possibly due to the use of MCD as the model selection criterion.

3.4.4 Visualizing the hidden representation space

In Section 3.2.1, it was hypothesized that applying the TTS-oriented pre-training technique results in an encoder capable of extracting linguistic-information-rich representation. To verify this hypothesis, a visualization experiment was conducted. Specifically,

1. The hidden representations were extracted with the trained encoders using the validation set from the *clb* speaker as input.
2. The t-SNE method [107] was used to reduce the dimensionality and visualize the hidden representations in the two-dimensional space.
3. The phoneme labels that come with the CMU ARCTIC dataset were used as ground truth, and the 5 most common phonemes and their corresponding hidden representations were colored to simplify the plots. Note that for encoders with a reduction factor greater than 1, the corresponding label was decided with majority voting. For example, if the encoder reduction factor was 4, and the labels of the four frames that correspond to a hidden representation are "s", "s", "s", "a", then the label of that hidden representation would be set to "s".

The resulting plots were shown in Figure 3.4. It could be observed that compared to no pre-training, the hidden representation spaces learned from TTS-oriented pre-training demonstrated a strong degree of clustering effect, where points corresponding to the same phoneme were close to each other. This tendency was consistent in the cases of both 932 and 80 training utterances. On the other hand, ASR-oriented pre-training yielded a much scattered hidden representation space even with 932 training utterances.

This analysis suggests that the TTS-oriented pre-training technique could result in a more discretized representation space, which matched the initial assumption. One may further conclude that by looking together with the objective and subjective results in Tables 3.1 and 3.2, the degree of clustering effect somehow reflected the goodness of the hidden representations for seq2seq VC.

3.5 Conclusions

In this chapter, a pre-training technique for addressing the problem of data efficiency in seq2seq VC was proposed. Specifically, a unified, two-stage training strategy that first pre-trains the decoder and the encoder subsequently followed by initializing the VC model with the pre-trained model parameters was proposed. ASR and TTS were chosen as source tasks to transfer knowledge from, and the RNN and VTN architectures were implemented. Through objective and subjective evaluations, it was shown that the TTS-oriented pre-training strategy could greatly improve performance in terms of speech intelligibility and quality when applied to both RNNs and VTNs, and the performance could stay without significant degradation even with limited training data. As for ASR-oriented pre-training, the robustness degraded with the reduction of training data size. Also, VTNs performed inferior to RNNs without pre-training but

superior with TTS-oriented pre-training. The visualization experiment suggested that the TTS-oriented pre-training could learn a linguistic-information-rich hidden representation space while the ASR-oriented pre-training lacks such ability, which sheds light on what an ideal hidden representation space would be like.

Table 3.1: *Validation-set objective evaluation results of VTNs with no pre-training, TTS-oriented pre-training, ASR-oriented pre-training, and RNN-based models with no pre-training and TTS-oriented pre-training, which are trained on different conversion pairs and different sizes of data. Bold font indicates the best performance across the average scores.*

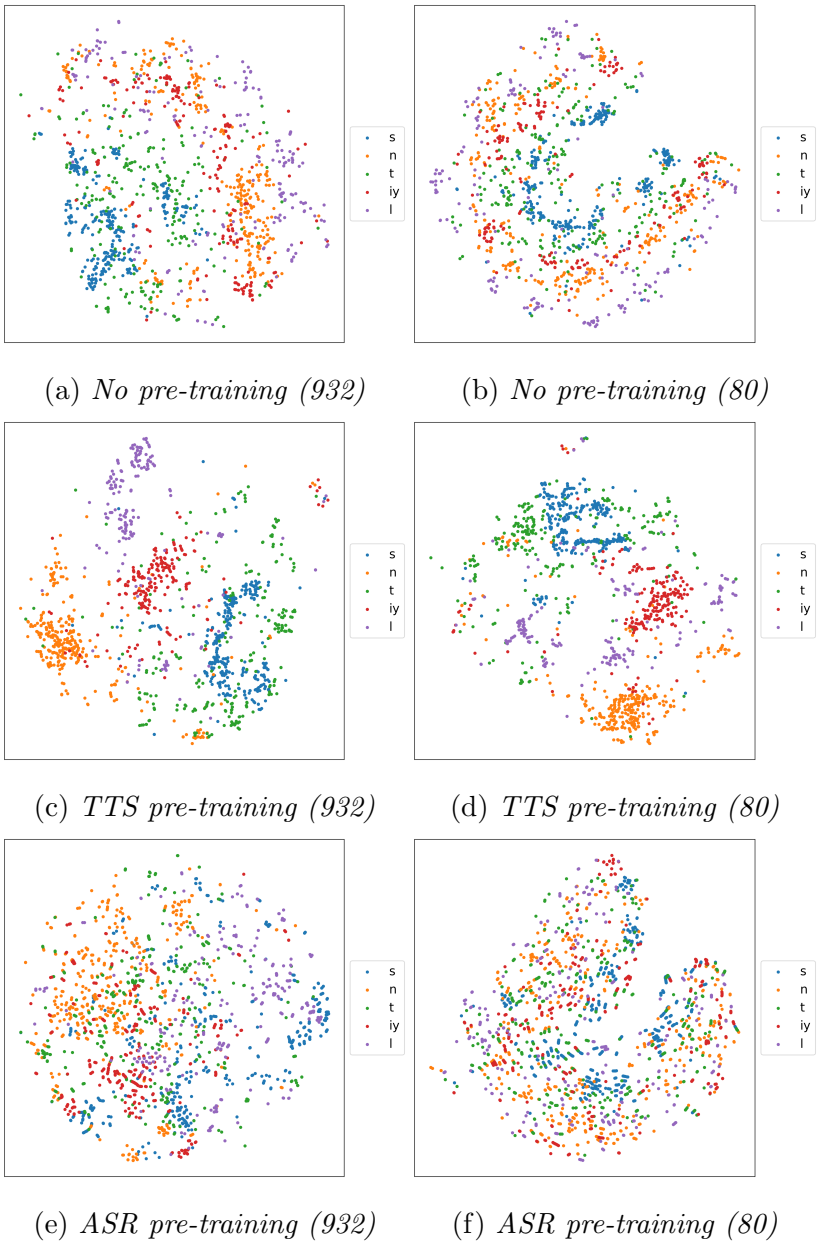
		Pair		932 training utterances				250 training utterances				80 training utterances			
Model	pre-train	Src	Trg	MCD	F0RMSE	CER	WER	MCD	F0RMSE	CER	WER	MCD	F0RMSE	CER	WER
VTN	None	clb	slt	6.60	24.54	12.4	20.2	7.43	25.37	29.2	42.3	8.23	26.65	65.3	87.6
			rms	6.83	24.02	21.4	33.0	7.83	24.86	53.2	73.3	8.68	27.20	71.8	94.8
		bdl	slt	7.33	23.36	23.1	33.7	8.31	23.40	52.9	75.9	8.74	24.81	73.9	95.7
			rms	7.37	22.68	28.4	43.3	8.30	23.78	56.2	78.4	9.14	24.38	79.0	102.8
		Average		7.03	23.65	21.3	32.6	7.97	24.35	47.9	67.5	8.70	25.76	72.5	95.2
	TTS	clb	slt	6.02	23.94	5.5	9.1	6.41	24.86	5.2	9.7	6.66	27.24	10.4	14.7
			rms	6.22	24.29	6.8	11.9	6.75	24.91	12.8	21.3	6.94	27.86	12.5	22.0
		bdl	slt	6.33	22.44	5.0	7.6	6.71	23.24	4.8	8.1	7.07	23.75	9.7	13.6
			rms	6.69	22.63	7.3	12.7	7.13	23.32	11.3	18.0	7.39	24.13	17.2	26.2
		Average		**6.32**	23.33	**6.2**	**10.3**	**6.75**	**24.08**	**8.5**	**14.3**	**7.02**	25.75	**12.5**	**19.1**
	ASR	clb	slt	6.11	24.03	4.8	10.9	6.84	24.78	15.9	26.0	8.28	27.13	72.1	97.6
			rms	6.22	24.15	8.1	16.0	7.08	24.89	27.2	43.2	7.93	26.57	60.2	86.2
		bdl	slt	6.50	22.35	5.7	11.1	7.33	23.65	26.1	39.8	8.18	24.23	58.2	80.7
			rms	6.68	22.46	9.1	15.6	7.58	23.24	32.9	51.6	8.22	24.25	59.7	82.9
		Average		6.38	**23.25**	6.9	13.4	7.21	24.14	25.5	40.2	8.15	25.55	62.6	86.9
RNN	None	clb	slt	6.77	24.81	7.1	12.1	7.29	25.02	15.4	24.0	7.76	25.04	38.6	56.8
			rms	6.80	23.54	11.6	19.7	7.49	24.84	24.7	38.0	7.98	27.67	48.9	68.7
		bdl	slt	7.45	23.37	23.4	32.6	8.06	24.53	37.1	54.4	8.44	24.40	65.6	93.8
			rms	7.62	23.96	20.0	32.4	8.25	24.32	47.2	90.2	8.52	25.13	59.7	81.5
		Average		7.16	23.92	15.5	24.2	7.77	24.68	31.1	51.7	8.18	25.56	53.2	75.2
	TTS	clb	slt	6.29	24.62	5.6	10.1	6.63	23.99	7.4	12.7	6.92	26.40	14.0	22.2
			rms	6.35	23.58	8.3	16.1	6.88	24.30	17.0	27.7	7.08	26.54	29.0	44.0
		bdl	slt	6.74	22.89	8.2	13.9	7.08	23.11	11.3	19.8	7.46	23.60	16.3	23.8
			rms	6.97	22.36	15.1	26.3	7.39	23.34	21.1	32.4	7.57	23.30	25.4	39.6
		Average		6.59	23.36	9.3	16.6	7.00	23.69	14.2	23.2	7.26	**24.96**	21.2	32.4

Table 3.2: *Evaluation-set naturalness and similarity subjective evaluation results of VTNs with no pre-training, TTS pre-training, ASR pre-training, and RNN-based models with no pre-training and TTS pre-training, which are averaged over all conversion pairs and different sizes of data.*

		932 training utterances		80 training utterances	
Model	pre-training	Naturalness	Similarity	Naturalness	Similarity
Analysis-synthesis		4.45 ± 0.14	-	-	-
VTN	None	3.19 ± 0.23	61% ± 14%	1.96 ± 0.16	44% ± 13%
	TTS	4.34 ± 0.15	80% ± 11%	4.11 ± 0.09	68% ± 8%
	ASR	4.25 ± 0.16	77% ± 10%	3.38 ± 0.20	53% ± 12%
RNN	None	2.33 ± 0.20	40% ± 12%	1.57 ± 0.14	33% ± 15%
	TTS	3.91 ± 0.19	68% ± 13%	3.71 ± 0.09	58% ± 10%

Table 3.3: *ASR-based recognition results of VTN converted samples from the evaluation set of the clb-slt conversion pair. The errors are in uppercase.*

Description	Training data size	Recognition result
Ground truth	-	the history of the eighteenth century is written ernest prompted
TTS-oriented pre-training	932	the history of the eighteenth century is written IN IS prompted TO TO
	250	the history of the eighteenth century is written IN HIS PROMPTER
	80	the history of the eighteenth century is written ON HIS PROMPT
ASR-oriented pre-training	932	the history of the eighteenth century is written EARNEST prompted
	250	the history of the eighteenth CENTURY'S RADIANCE prompted
	80	IT DISTURBED the DAY TO HIMSELF TO REJOIN HIM IN NORTH'S LIBRARY



(a) *No pre-training (932)* (b) *No pre-training (80)*

(c) *TTS pre-training (932)* (d) *TTS pre-training (80)*

(e) *ASR pre-training (932)* (f) *ASR pre-training (80)*

Figure 3.4: *Visualizations of hidden representations extracted from VTNs with no pre-training, TTS-oriented pre-training, and ASR-oriented pre-training. The validation set from clb was used. The numbers in the parenthesis indicate the number of training utterances.*

4 Self-supervised Pre-training for Voice Conversion

In this chapter, self-supervised pre-training-based voice conversion (VC) is studied. The study presented in this chapter is part of a series of scientific activities, which is an extension of the self-supervised speech pre-training and representation learning (S3PRL) toolkit and the Speech processing Universal PERformance Benchmark (SUPERB). The study carried out in this chapter is a collection of toolkit, benchmark, and experimental results, which is referred to as **S3PRL-VC**. With the success of self-supervised learning (SSL) in other research fields, its application to VC is highly anticipated.

4.1 Introduction

In recent years, SSL has become the state-of-the-art approach in various research fields. It implies a principle that first pre-trains an *upstream* model that learns general knowledge by solving self-supervised tasks on a large amount of unlabeled data, followed by fine-tuning prediction layers on various *downstream* tasks[1]. An upstream model pre-trained for speech is called a self-supervised speech representation (S3R) model. These

[1]In the context of SSL-based VC, the recognizer is represented and the synthesizer are represented by the self-supervised speech representation upstream model and the downstream prediction layers, respectively. In the remainder of this paper, we will use these two terms interchangeably.

S3Rs are expected to capture linguistic, speaker, prosodic, and semantic information of speech. In the literature, though with different network architectures, S3Rs are commonly grouped by their objective functions.

- **Generative** modeling incorporates language model-like training losses to predict unseen regions (such as future or masked frames) to maximize the likelihood of the observed data. Examples include APC [108], VQ-APC [109], Mockingjay [110], TERA [111], and NPC [112].

- **Discriminative** modeling aims to discriminate (or contrast) the target unseen frame with randomly sampled ones, which is equivalent to mutual information maximization. Examples include CPC [113,114], wav2vec [115], vq-wav2vec [116], wav2vec 2.0 [38] and HuBERT [39].

- **Multi-task learning** applies multiple objectives, including waveform generation, prosody features regression, and contrastive InfoMax. PASE+ [117] is the most representative approach.

As introduced in Section 2.2.2, one can use an S3R model as the recognizer in the context of recognition-synthesis (rec-syn)-based VC. Such a framework is referred to as S3R-based VC in this thesis. S3R-based VC can be especially attractive due to its unique advantages. The first advantage is that it unlocks the amount of training data that can be used to train the recognizer. For instance, the 960 hours LibriSpeech dataset [100] is often used to train an automatic speech recognition (ASR) model as the recognizer in the rec-syn-based VC systems based on text or phonetic posteriorgram (PPG), as described in Section 2.2.2. In contrast, most S3Rs are benchmarked on the 60k hours LibriLight dataset [118], which is 60 times larger than LibriSpeech. Another advantage is that it is much easier to collect unlabeled data for atypical speech types,

such as emotional, accented, or low-resourced languages.

Perhaps the most unique advantage of applying S3R to VC is its use of analyzing S3R models. Take speaker conversion as an example, from the information perspective of VC presented in Section 2.2, it is hypothesized that a good intermediate representation should (1) be rich in content but (2) contain little to none speaker information. As a result, an S3R model that can extract all-purpose speech representations may not be an optimal choice for VC. For instance, a well-known S3R, wav2vec 2.0 [38], is powerful in not only ASR but also speaker and language recognition [119], implying that it encodes rich content, speaker and language information. Therefore, it may not be the best representation for VC. Such analyses may help researchers reach a better understanding of different S3R models.

In this chapter, a comparative study of S3R-based VC is presented. The presented results are the outcome of the development of S3PRL-VC, [120], an open-source VC software[2] that was previously developed to extend the SUPERB benchmark and the S3PRL toolkit [121]. A large-scale evaluation, both objectively and subjectively, is conducted to analyze S3R-based VC systems from various aspects, including:

- **Task.** Experiments are conducted under three kinds of settings: intra-/cross-lingual any-to-one (A2O) VC, where the system converts from an unseen speaker to a seen speaker of the same/different language, and intra-lingual any-to-any (A2A) VC, where both the source and target speakers are unknown during training. The voice conversion challenge (VCC) 2020 dataset is used to unify the dataset condition and to provide a comparison with the top systems in the challenge.

[2]Originally the implementation was released as a part of S3PRL: `https://github.com/s3prl/s3prl/tree/master/s3prl/downstream/a2o-vc-vcc2020`. Later on, an isolated toolkit was released at `https://github.com/unilight/s3prl-vc`.

- **Model type.** Models used in the top systems in VCC2018 [89] and VCC2020 [78] are implemented, which allows comparison with the top systems in the respective years.

- **Multilinguality.** The cross-lingual transfer ability of S3Rs is validated using a cross-lingual VC task. Furthermore, using the wav2vec 2.0 model, performance when trained on a mono-lingual and a multi-lingual dataset is compared.

- **Supervision.** Results of supervised representations-based systems using the same tasks and models are presented to understand the impact of supervision in recognizer training.

- **Discretization.** Although continuous features were used as default in the SUPERB benchmark, initial investigations showed that they do not provide the sufficient disentanglement needed in the A2A setting. Therefore, the use of a discretization technique based on the k-means clustering algorithm used in [81] is studied, and an extensive is carried out.

This chapter aims to contribute to not only the VC field but also the S3R field. The contributions to the respective fields are summarized as follows:

- **VC**: The goal is a unified, comprehensive study of S3R-based VC. Although getting increasingly popular in the VC field in recent years [81–85], each paper used its own experimental setting, including different datasets, models, and evaluation protocol. As a result, it is difficult to compare different techniques to further identify the drawbacks of current methods. Through this study, it is expected that a more holistic understanding of the S3R-based VC framework can be delivered, to provide a stepping stone for future VC researchers.

Table 4.1: *Summary of the data conditions in the voice conversion challenge 2020.*

Task	Training phase		Conversion phase	
	Source	Target	Source	Converted
Task 1	70 Eng. utterances	70 Eng. utterances	25 Eng. utterances	25 Eng. utterances
Task 2		70 Man./Ger./Fin. utterances		

- **S3R**: VC is expected to be a suitable task for investigating the disentanglement ability of S3R models. Most downstream tasks test one ability of the S3R model at a time, either the capability to encode rich and compact local content information (speech recognition, keyword spotting, etc.) or the power to represent global characteristics (speaker verification, emotion recognition, etc.) As stated above, it is expected that VC can test these two abilities at once. Moreover, although speaker conversion is the main focus of this work, by changing a task setting, it is possible to inspect the ability of the S3R model to disentangle different global attributes, such as accent or speaking style.

4.2 Tasks Design

4.2.1 General description of the voice conversion challenge 2020 dataset and tasks

All experiments in this work are based on the VCC2020 dataset [9]. There are two tasks in VCC2020, both of which are speaker conversion tasks: task 1 is intra-lingual

VC, and task 2 is cross-lingual VC. The data conditions are summarized in Table 4.1. The two tasks share the same two English male and female source speakers. The target speakers include two male and two female English speakers for task 1, and one male and one female speaker each of Finnish, German, and Mandarin for task 2. For each speaker, 70 utterances (roughly five minutes) in their respective languages and contents are provided, and there are 25 test sentences for evaluation. During conversion, the source speech (which is in English) is converted as if it was uttered by the target speaker while keeping the linguistic contents unchanged.

4.2.2 Intra-lingual and cross-lingual any-to-one VC

The first two tasks considered in this study are intra-lingual and cross-lingual A2O VC. A2O VC refers to the task of converting any arbitrary speech into that of a pre-defined target speaker. Such a definition naturally makes rec-syn-based VC a suitable approach to this task. In the A2O setting, the training utterances of the source speakers in Table 4.1 are not used, and only the target training utterances are used as the VC training dataset in Figure 2.6. As described in Secion 4.2.1, the language of the target training dataset is English and Finnish/German/Mandarin in the intra-lingual and cross-lingual settings, respectively.

A2O VC is a good probing task to investigate several characteristics of an upstream S3R model. A fundamental requirement of VC is linguistic consistency, so it is believed that there is a positive correlation between the VC performance of an S3R model and its ability to faithfully encode the spoken contents. Also, during the synthesizer training in cross-lingual VC, the S3R model may fail to generalize to non-English target speakers since most existing S3R models are trained with English datasets only. It is worthwhile to examine the ability of mono-lingual S3R models to transfer to different languages.

4.2.3 Intra-lingual any-to-any VC

The third task is A2A VC with the VCC2020 dataset. A2A VC is also known as one-shot VC[3], which attempts to convert to a target speaker whose training set is so limited (often less than one minute), such that fine-tuning is infeasible.

Certain modifications need to be made to adapt the training framework illustrated in Figure 2.6 for A2A VC. Since S3Rs are speaker-independent, they cannot provide the essential information to recover the speaker identity for reconstruction. Thus, in the A2A setting, the input to the synthesizer is augmented with a speaker embedding extracted by an off-the-shelf speaker encoder, which is pre-trained on an automatic speaker verification (ASV) dataset and objective. In training, the speaker embedding extracted from the target waveform is used. During conversion, given a small set of utterances of the target speaker, the speaker embedding is formed as an average of each embedding from each utterance. We may then rewrite Eq. 2.17 as:

$$\mathbf{Y} = \text{Synth}(\mathbf{H}, \mathbf{s}), \mathbf{H} = \text{Recog}(\mathbf{X}), \mathbf{s} = \text{SpkEnc}(\mathbf{D}_{\text{trg}}), \tag{4.1}$$

where $\mathbf{s}$ is the speaker embedding. In practice, a separate multi-speaker dataset is used to train the synthesizer, such that it learns to generalize to new speakers at test time.

A2A VC is considered more difficult than A2O VC. The reason is that during synthesizer training, the speaker identity that the model tries to reconstruct is drawn randomly from the multi-speaker dataset, instead of always reconstructing the same target speaker as in the training of a synthesizer in A2O VC. In such a scenario, a

[3]The term "zero-shot VC" was recently used in some widely cited VC papers [63,122]. However, the term "zero-shot" has been mostly used in machine learning and refers to the ability to adapt to unseen *classes* in discriminative tasks. It is thus questionable whether such a term matches the condition in generative tasks since it is impossible to generate a certain class without knowing anything about that class. For instance, if one wants to synthesize the voice of a certain speaker, certain information must be given. The author finds "one-shot" to be a more suitable choice, as in [70].

"speaker-information-free" S3R will be more demanding.

4.3 Implementations

4.3.1 Recognizers (upstream models)

Table 4.2 depicts the list of S3Rs we compared in this work, which are the upstream models supported in S3PRL as of August 2021. For a complete list of information (training data, architecture, objective, etc.), please refer to [121]. All upstreams are trained with English data, mostly LibriSpeech [100] or LibriLight [118]. In addition to the S3Rs, two extra upstreams were included: (1) mel-spectrogram, "mel", and (2) "PPG (TIMIT)", which is trained supervisedly on the TIMIT dataset [123].

4.3.2 Synthesizer models

As illustrated in Figure 4.1, the following models are implemented to resemble top systems of past VCCs:

- **Simple**: The first model resembles the top system in VCC2018 [89]. The simple model consists of a single-layer feed-forward network (FFN), two long short-term memory layers with projection (LSTMP), and a linear projection layer.
- **Simple-AR**: As autoregressive (AR) modeling has been shown to be effective in speech synthesis [124], an AR loop is added to the simple model. At each time step, the previous output is consumed by the first LSTMP layer. Dropout is essential in the AR loop to avoid exposure bias brought by teacher-forcing [41, 125].

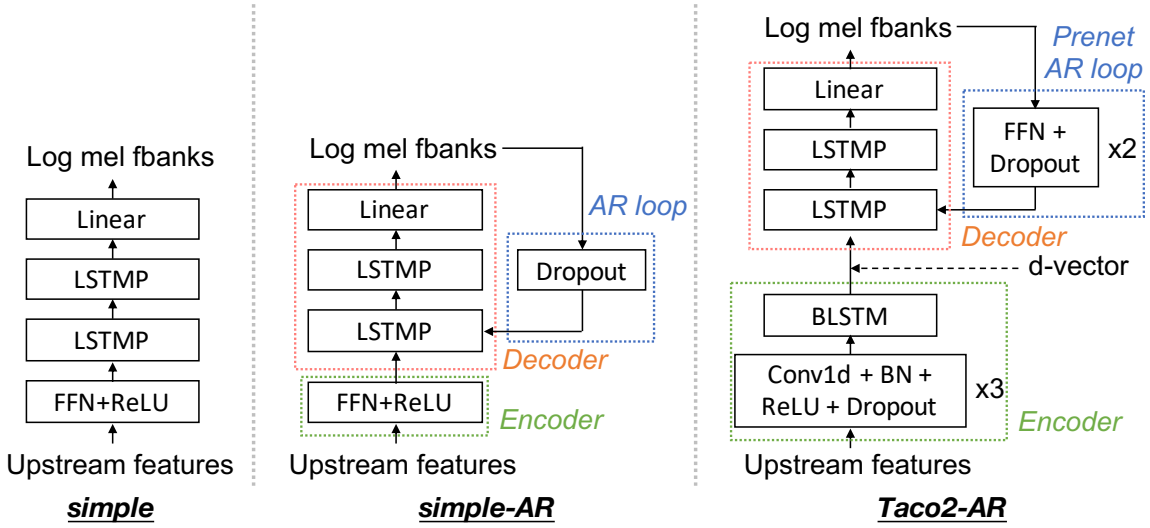


Figure 4.1: *The synthesizer models implemented in this study. Left: the simple model. Middle: the simple model with an auto-regressive loop. Right: the Tacotron2 model, with extension to an any-to-any model by accepting a d-vector as the speaker embedding.*

- **Taco2-AR**: Finally, a model architecture similar to that of Tacotron 2 [42] is implemented, which resembles the model used by the top system in VCC2020 [78]. Different from Tacotron 2, the attention module was not used as it was reported to be not necessary [78].

The transformer architecture is not used because (1) fast benchmarking is a key requirement of SUPERB/S3PRL, and (2) using the frame-level feature used in the S3R-based framework does not require changing the temporal structure. To unify the acoustic feature to reconstruct, the log mel-spectrogram is used. The input of the synthesizer can be either the raw continuous S3R, or discretized with embedded vectors as will be described in Section 4.3.3.

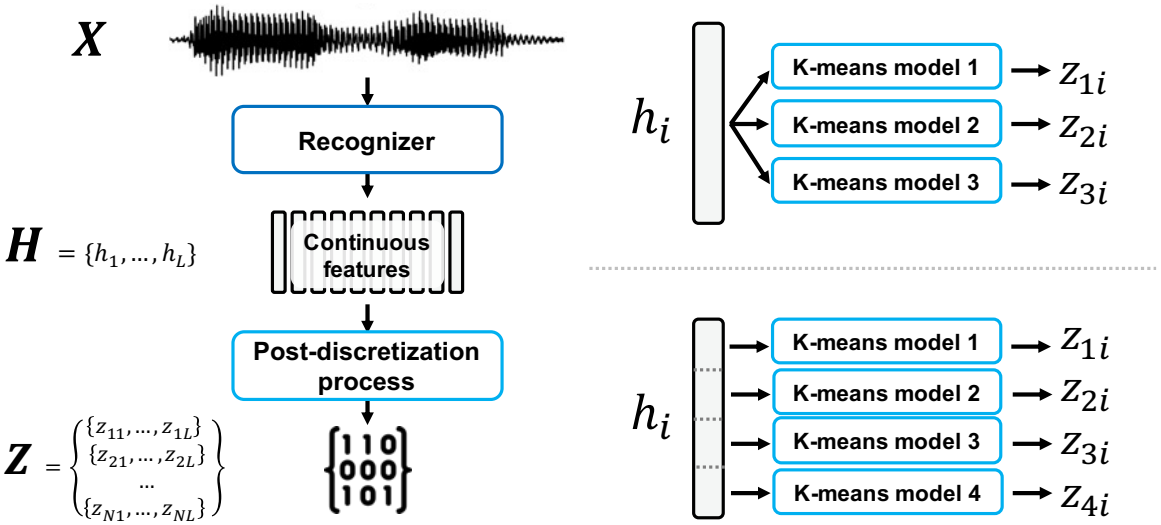


Figure 4.2: *Left: the post-discretization process overview. Top right: the cluster ensemble technique with three k-means models using different numbers of clusters. Bottom right: the product quantization techniques with four partitions.*

4.3.3 Post-discretization process for any-to-any VC

As will be discussed later (and shown in Table 4.2), using continuous features cannot satisfy the disentanglement requirement in the A2A scenario. As a result, most S3Rs fail to convert the speaker identity, as will be shown in later sections. This makes A2A not a suitable task for benchmarking S3Rs since most of the S3Rs perform similarly badly.

In light of this, the post-discretization process proposed in [81] is adopted with the motivation to impose a stronger information bottleneck. Specifically, as illustrated in the left of Figure 4.2, the k-means clustering algorithm takes the continuous features returned by the recognizer, and returns corresponding discrete indices $\mathbf{Z}$ using a codebook of size K. The k-means model is trained with a separate dataset in advance.

However, such a discretization technique performs poorly when naively applied to

certain S3Rs. The generated speech often suffers from poor intelligibility, even when using a large codebook. It is suspected that the information bottleneck introduced by discretization is too strong. To offer more expressive power, inspired by [39], the following two additional techniques are applied. The right of Figure 4.2 shows such a process. In particular, both methods try to describe one feature vector with multiple k-means models (i.e., multiple indices) to increase the degree of freedom. In the experimental section, a complete investigation of these two techniques will be presented.

Cluster ensemble

Using an ensemble of k-means models with different codebook sizes can capture different granularity, and each k-means model can provide complementary information to back each other up. Specifically, given h_i, a continuous feature vector, N_{CE} k-means models are used to generate N_{CE} indices: $[z_{1i}, z_{2i}, \ldots, z_{N_{\text{CE}}i}]$, where the codebook of n-th model has size K_n clusters. Each K_n should be set to different numbers so that different k-means models can learn to capture different levels of detail. The hyperparameter N_{CE} can be tuned for each S3R, and an ablation study on this hyperparameter will be presented in Section 4.4.7

Product quantization

Product quantization (PQ) is a technique where the feature space is partitioned into multiple subspaces, and each subspace is quantized separately using different k-means models. Specifically, given a continuous feature vector $h_i \in \mathbb{R}^d$, it is first partitioned into N_{PQ} subvectors: $[h_{1i}, h_{2i}, \ldots, h_{N_{\text{PQ}}i}]$ where each subvector has size $h_{ni} \in \mathbb{R}^{d/N_{\text{PQ}}}$. Then, each subvector is consumed by a separate k-means model to generate N_{PQ} indices: $[z_{1i}, z_{2i}, \ldots, z_{N_{\text{PQ}}i}]$. The k-means models can be of different numbers of clusters

as done in cluster ensemble, but for simplicity, here all k-means models are set to have an equal number of clusters. Similar to N_{CE}, one can also tune N_{PQ} for each S3R, and an ablation study on this hyperparameter will be presented in Section 4.4.7

4.3.4 Other implementation setups

Any-to-any VC settings

The dataset used to train the A2A VC model is the VCTK dataset [11]. For the speaker encoder, the d-vector model [95] is used and trained on a mix of datasets, including LibriSpeech, VoxCeleb 1 [126] and 2 [127]. For the post-discretization process, following [81], all k-means models are trained on the LibriSpeech clean-100h set [100]. Although some studies use intermediate layer outputs for discretization [81, 128], for simplicity, the last output for all S3R models is used.

Waveform synthesizer

HiFi-GAN [129], a state-of-the-art parallel real-time neural vocoder, is used to generate the final waveform from the generated mel spectrogram. For the A2O setup, the training utterances of all 14 speakers in VCC2020 and the VCTK dataset are used, while for the A2A setup, only the VCTK dataset is used.

4.4 Experimental evaluation results

In this section, a series of complete objective evaluations and a large-scale listening test are reported to analyze *continuous* feature-based S3R-based VC and to compare with state-of-the-art systems (Section 4.4.4). The investigated aspects include the

synthesizer model type (Section 4.4.2), multilinguality (Section 4.4.3), and supervision (Section 4.4.5). Finally, the effectiveness of the post-discretization process is examined (Sections 4.4.7 and 4.4.8).

4.4.1 Evaluation metrics and protocols

For objective measures, the MCD, WER, and ASV metrics described in Section 2.3 are reported. The ASR engine for calculating WER is a pretrained wav2vec 2.0 model[4]. For the cross-lingual A2O task, the MCD numbers are not reported due to the absence of ground truth reference.

For the subjective test, naturalness and similarity are evaluated, as described in Section 2.3. 80 utterances (5 random $\times$ 16 conversion pairs) were evaluated for each system. Recordings of the target speakers were also included in the naturalness test and served as the upper bound. An open-source toolkit [105] that implemented the ITU-T Recommendation P.808 [106] was used to screen unreliable ratings obtained through Amazon Mechanical Turk (Mturk). More than 280 listeners from the United States were recruited, and each sample was rated by five different participants on average. Audio samples are available online[5].

4.4.2 Comparison of different synthesizer model types

As a start, the impact of using different synthesizer models described in Section 4.3.2 in the intra-lingual A2O setting is investigated, as shown in Table 4.2. First, only by adding the AR loop to the Simple model, most S3Rs benefit from large improvements

[4]Performance and APIs can be found at `https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self`

[5]`https://unilight.github.io/Publication-Demos/publications/s3prl-vc/`

Table 4.2: *Objective evaluation results on **intra-lingual any-to-one** voice conversion over various **continuous** self-supervised speech representations. For MCD and WER, the smaller the better; for ASV, the higher the better.*

Upstream	Intra-lingual A2O								
	Simple			Simple-AR			Taco2-AR		
	MCD	WER	ASV	MCD	WER	ASV	MCD	WER	ASV
mel	8.41	48.5	59.00	8.92	22.7	49.75	8.47	38.3	77.25
PPG (TIMIT)	7.78	69.0	85.50	7.83	58.9	95.25	7.18	33.6	99.75
PASE+ [117]	9.29	5.0	26.75	9.52	5.7	26.00	8.66	30.6	63.20
APC [108]	8.67	8.6	48.00	8.73	7.1	41.75	8.05	27.2	87.25
VQ-APC [109]	8.12	10.8	81.25	8.37	7.4	60.50	7.84	22.4	94.25
NPC [108]	7.74	39.0	92.75	8.15	21.1	76.75	7.86	30.4	94.75
Mockingjay [110]	8.58	31.3	51.00	8.74	9.5	47.00	8.29	35.1	79.75
TERA [111]	8.60	11.4	46.50	8.67	6.0	42.50	8.21	25.1	83.75
Modified CPC [114]	8.71	9.4	40.00	8.87	7.0	30.00	8.41	26.2	71.00
DeCoAR 2.0 [130]	8.31	7.4	54.75	8.33	6.4	53.00	7.83	17.1	90.75
wav2vec [115]	7.45	14.0	**95.50**	7.64	4.9	90.50	7.45	10.1	98.25
vq-wav2vec [116]	**7.41**	13.4	91.00	**7.24**	11.6	**98.75**	**7.08**	13.4	**100.00**
wav2vec 2.0 Base [38]	7.80	24.7	92.75	7.77	5.0	86.50	7.50	10.5	98.00
wav2vec 2.0 Large	7.64	12.5	81.75	7.67	9.0	82.75	7.63	15.8	97.25
HuBERT Base [39]	7.70	**5.5**	89.25	7.79	**4.7**	84.25	7.47	**8.0**	98.50
HuBERT Large	7.54	5.6	95.00	7.54	5.6	93.00	7.22	9.0	99.25

in WER. With Taco2-AR, all S3Rs except PASE+ and modified CPC achieved an ASV accept rate higher than 80%, while all S3Rs suffered from a degradation in WER. This shows that increasing the model capacity can significantly improve speaker similarity while sacrificing intelligibility.

However, it should be noted that: (1) the WER is a strict measurement of intelligibility, and humans can recognize better than machines. After listening to the samples, the internal perception was that compared to simple-AR, the quality was greatly im-

proved, and intelligibility degradation was not as serious as shown in the table. (2) the Taco2-AR model yields the best MCD scores, which, as will be shown later, correlates better with subjective naturalness and similarity. (3) although the Taco2-AR is considered the most complicated, it was found empirically that the training times of the three models are similar. Based on these reasons, the Taco2-AR model is used for the succeeding tasks and comparisons.

Table 4.3: *Objective evaluation results on* ***cross-lingual any-to-one*** *and* ***intra-lingual any-to-any*** *voice conversion over various* ***continuous*** *self-supervised speech representations. For MCD and WER, the smaller the better; for ASV, the higher the better.*

Upstream	Cross-lingual A2O		Intra-lingual A2A		
	Taco2-AR		Taco2-AR		
	WER	ASV	MCD	WER	ASV
mel	39.0	46.67	9.49	4.2	19.50
PPG (TIMIT)	51.0	84.67	**8.31**	12.9	**83.50**
PASE+ [117]	36.3	34.67	9.85	4.2	8.00
APC [108]	33.9	52.33	9.57	3.5	23.25
VQ-APC [109]	28.4	68.00	9.43	4.0	22.00
NPC [108]	37.6	59.00	9.39	4.4	21.00
Mockingjay [110]	39.2	46.00	9.43	5.0	25.00
TERA [111]	29.2	49.33	9.31	5.2	18.75
Modified CPC [114]	35.3	32.83	9.61	4.1	10.75
DeCoAR 2.0 [130]	26.8	59.33	9.28	4.0	27.00
wav2vec [115]	13.9	75.83	8.77	3.5	40.00
vq-wav2vec [116]	21.0	**88.83**	8.47	4.2	73.25
wav2vec 2.0 Base [38]	14.9	82.17	9.03	3.2	27.00
wav2vec 2.0 Large	22.7	78.00	8.99	4.1	22.25
HuBERT Base [39]	**13.5**	82.33	9.19	3.4	23.25
HuBERT Large	15.9	86.50	9.13	**3.0**	27.75

Table 4.4: *Comparison of wav2vec 2.0 trained on mono-lingual data and cross-lingual data in the cross-lingual any-to-one voice conversion scenario, using the Taco2-AR model. The results of wav2vec 2.0 Large are extracted from Table 4.3.*

Upstream	Training data size	WER	ASV
wav2vec 2.0 Large	LibriLight 60k hr	22.7	78.00
XLSR [131]	56k hr from 53 languages	24.2	72.50

4.4.3 Investigation on model multilinguality

Next, Table 4.3 shows the VC performance of S3R models in the cross-lingual setting. First, S3Rs trained on a mono-lingual corpus can still work well in the cross-lingual setting, demonstrating their ability to transfer across languages. However, compared with the intra-lingual A2O task, it could be observed that all S3Rs degraded in terms of both the WER and ASV accept rate in the cross-lingual setting. In VCC2020, it was also reported that cross-lingual VC is indeed a harder task than intra-lingual VC, as the listening test results of all participating teams were much worse.

To further investigate the impact of the training data language, results are carried out on XLSR [131], a model that has the same architecture as wav2vec 2.0 Large but trained on a mixture of datasets from 53 languages, resulting in 56k hours of data. From Table 4.4, it can be found that compared to wav2vec 2.0 Large trained on mono-lingual data, XLSR was not particularly good. One potential reason is that when the training set is large enough, the model can already capture the variations among all languages such that a multilingual dataset will not be needed. Also, since the source language during conversion is English, monolingual models may be sufficient. It is worthwhile investigating this point by considering a different setting in the future, such as converting from non-English languages.

Table 4.5: *Comparison of S3R-based systems and state-of-the-art systems in the **any-to-one** setting. All upstreams use the Taco2-AR model. Nat. and Sim. stand for naturalness and similarity, respectively. Both Nat. and Sim. are the higher the better. The objective results (MCD, WER, ASV) are extracted from Table 4.2.*

System	MCD	WER	ASV	Nat.	Sim.
Intra-lingual A2O					
mel	8.47	38.3	77.25	2.61 ± .11	35% ± 3%
PPG (TIMIT)	7.18	33.6	99.75	3.32 ± .10	58% ± 4%
PASE+	8.66	30.6	63.20	2.58 ± .12	31% ± 3%
APC	8.05	27.2	87.25	2.92 ± .11	43% ± 4%
VQ-APC	7.84	22.4	94.25	3.08 ± .10	40% ± 4%
NPC	7.86	30.4	94.75	2.98 ± .11	46% ± 3%
Mockingjay	8.29	35.1	79.75	2.81 ± .12	42% ± 4%
TERA	8.21	25.1	83.75	2.91 ± .12	37% ± 4%
Modified CPC	8.41	26.2	71.00	2.74 ± .11	33% ± 3%
DeCoAR 2.0	7.83	17.1	90.75	3.04 ± .11	43% ± 4%
wav2vec	7.45	10.1	98.25	3.40 ± .05	52% ± 2%
vq-wav2vec	7.08	13.4	100.00	3.59 ± .10	59% ± 4%
wav2vec 2.0 B.	7.50	10.5	98.00	3.36 ± .06	51% ± 2%
wav2vec 2.0 L.	7.63	15.8	97.25	3.26 ± .10	50% ± 4%
HuBERT B.	7.47	8.0	98.50	3.48 ± .10	55% ± 4%
HuBERT L.	7.22	9.0	99.25	3.47 ± .10	54% ± 4%
USTC-2018† [89]	–	6.5	99.00	4.20 ± .08	55% ± 4%
USTC-2020 [77]	6.98	5.4	100.00	4.41 ± .07	82% ± 3%
SRCB [80]	8.90	11.5	92.00	4.16 ± .08	68% ± 3%
CASIA [79]	7.13	11.0	98.25	4.25 ± .08	61% ± 4%
ASR+TTS [76]	6.48	8.2	100.00	3.84 ± .09	75% ± 3%
Target	–	0.7	–	4.57 ± 0.14	–

4.4.4 Comparing with state-of-the-art systems using subjective evaluation

This subsection presents a comparison of S3R-based VC models with state-of-the-art VC systems in VCC2020. **USTC-2018** [89], **USTC-2020** [77, 78][6], **SRCB** [80], **CASIA** [79] were top systems in VCC2020, all of which adopted PPGs, synthesizer

[6]USTC's systems used text and PPG for the intra-lingual and cross-lingual tasks, respectively.

Table 4.6: *Comparison of S3R-based systems and state-of-the-art systems in the **cross lingual any-to-one** and **intra-lingual any-to-any** settings. All upstreams use the Taco2-AR model. Nat. and Sim. stand for naturalness and similarity, respectively. Both Nat. and Sim. are the higher the better. The objective results (MCD, WER, ASV) are extracted from Table 4.3.*

System	MCD	WER	ASV	Nat.	Sim.
Cross-lingual A2O					
PPG (TIMIT)	–	51.0	84.67	2.79 ± .08	43% ± 3%
vq-wav2vec	–	21.0	88.83	3.28 ± .08	44% ± 3%
HuBERT L.	–	15.9	86.50	3.13 ± .08	41% ± 3%
USTC-2018 [89]	–	5.6	97.67	4.17 ± .06	34% ± 3%
USTC-2020 [78]	–	7.6	96.00	4.27 ± .07	43% ± 3%
SRCB [80]	–	8.6	78.67	4.34 ± .07	34% ± 3%
CASIA [79]	–	10.5	91.67	4.11 ± .07	45% ± 3%
ASR+TTS [76]	–	34.5	67.83	2.51 ± .08	39% ± 3%
Target	–	–	–	4.48 ± 0.12	–
Intra-lingual A2A					
PPG (TIMIT)	8.32	12.7	84.25	3.41 ± .08	34% ± 4%
vq-wav2vec	8.47	4.2	73.25	3.58 ± .09	28% ± 3%
S2VC† [84]	–	12.4	71.50	2.90 ± .09	29% ± 3%

†: Systems generate 16kHz, so MCD is not calculable and direct score comparison should be made with caution.

pretraining on a multi-speaker dataset, and AR vocoders. Notably, they used thousands of hours of internal data for training. **ASR+TTS** [76] was the seq2seq+non-AR vocoder baseline in VCC2020. **S2VC** [84] is the STOA system for A2A VC.

The results are shown in Tables 4.5 and 4.6, and a summary of the observations are as follows:

- vq-wav2vec outperformed all other upstreams in the subjective test, with a 3.59 naturalness and 59% similarity in the intra-lingual A2O setting.
- In the A2O settings, there was still a naturalness gap between vq-wav2vec and

other VCC2020 top systems (3.59 v.s. 4.16-4.25, 3.28 v.s. 4.11-4.34). As for similarity, vq-wav2vec was on par with USTC-2018 and CASIA in the intra-lingual A2O setting and achieved top in the cross-lingual setting.

- In the A2A setting, vq-wav2vec was on par with S2VC in similarity, while being significantly better in naturalness. The system presented in this chapter is therefore the new state-of-the-art in S3R-based A2A VC.

4.4.5 Impact of supervision

Although top systems using PPG greatly outperformed vq-wav2vec in naturalness, they used AR vocoders, and the system was trained on large internal datasets, so the impact of supervision is not yet clear. To this end, a comparison is made with the vq-wav2vec result with "PPG (TIMIT)" and the same vocoder.

From Tables 4.5 and 4.6, it is first found that "PPG (TIMIT)" has a high WER and a low naturalness score, showing that it was indeed of low quality. Nonetheless, in all three settings, "PPG (TIMIT)" can achieve similar or higher similarity scores than vq-wav2vec. This shows that supervision greatly contributes to similarity, especially in a difficult setting like A2A VC. This also shows that the ability of current S3Rs to disentangle speaker information is still limited when compared to PPG, and can be further improved in the future. That being said, good performance can still be achieved without supervision if the S3R is designed properly.

4.4.6 Justify the objective metrics with correlation analysis

Conducting a subjective test whenever a new S3R is developed cannot meet the fast benchmark requirement of SUPERB. Therefore, it is necessary to examine if the

Table 4.7: *Linear correlation coefficients between different metrics.*

Metric	MCD	WER	ASV	Nat.	Sim.
MCD	–	0.678	-0.934	-0.968	-0.961
WER	–	–	-0.640	-0.808	-0.587
ASV	–	–	–	0.910	0.911
Nat.	–	–	–	–	0.932
Sim.	–	–	–	–	–

objective measures align well with human perception. Table 4.7 shows the pairwise linear correlation coefficients calculated using the intra-lingual A2O results over different upstream. It could be found that MCD is more aligned with the subjective scores in terms of both naturalness and similarity than all other metrics.

Note that in this correlation analysis, all systems used the same synthesizer and neural vocoder. Since the correlation result is strongly affected by the pool of methods evaluated in a listening test, this good correlation could be observed only in such a homogeneous condition. That is to say, as long as the synthesizer and the vocoder are the same, it is safe to use the objective measures to compare different upstreams. This implication is very useful for the benchmarking requirement of SUPERB.

4.4.7 Investigation of the post-discretization process

As mentioned in Section 4.3.3, the results in Table 4.3 suggest that using *continuous* S3Rs in the A2A setting makes it difficult to properly evaluate the S3Rs, since all S3R models performed badly. Therefore, in this subsection, the focus is to examine whether the discretization process described in Section 4.2 can alleviate this problem.

Table 4.8: *Results of HuBERT Base and Mockingjay using the cluster ensemble and the product quantization techniques in the any-to-any scenario, with the Taco2-AR model. The best numbers within the same upstream are in boldface.*

Upstream	# clusters (K_n)	N_{PQ}	MCD	WER	ASV
HuBERT Base	50	1	8.41	22.0	79.50
	100		8.25	10.3	83.50
	200		8.32	10.2	84.25
	50+100		8.37	10.2	**86.25**
	50+200		8.28	8.6	84.25
	100+200		8.29	8.8	83.25
	50+100+200		8.40	7.9	85.00
	50	2	8.37	12.7	84.50
	100		**8.23**	8.2	**86.25**
	200		8.32	**7.2**	81.75
Mockingjay	100	1	9.12	77.4	63.00
	200		9.10	73.1	**63.25**
	50+100+200		9.02	59.7	61.00
	100	2	9.07	64.5	59.00
	200		**8.95**	**55.4**	61.75

Table 4.8 reports the results of applying cluster ensemble and PQ on two upstreams, namely HuBERT Base and Mockingjay, in the A2A setting. First, the intelligibility (WER) improves when the number of k-means models in the ensemble increases. That is to say, using two k-means models is better than using one, and using three is even better. The intelligibility is also improved when using PQ, and the improvement is

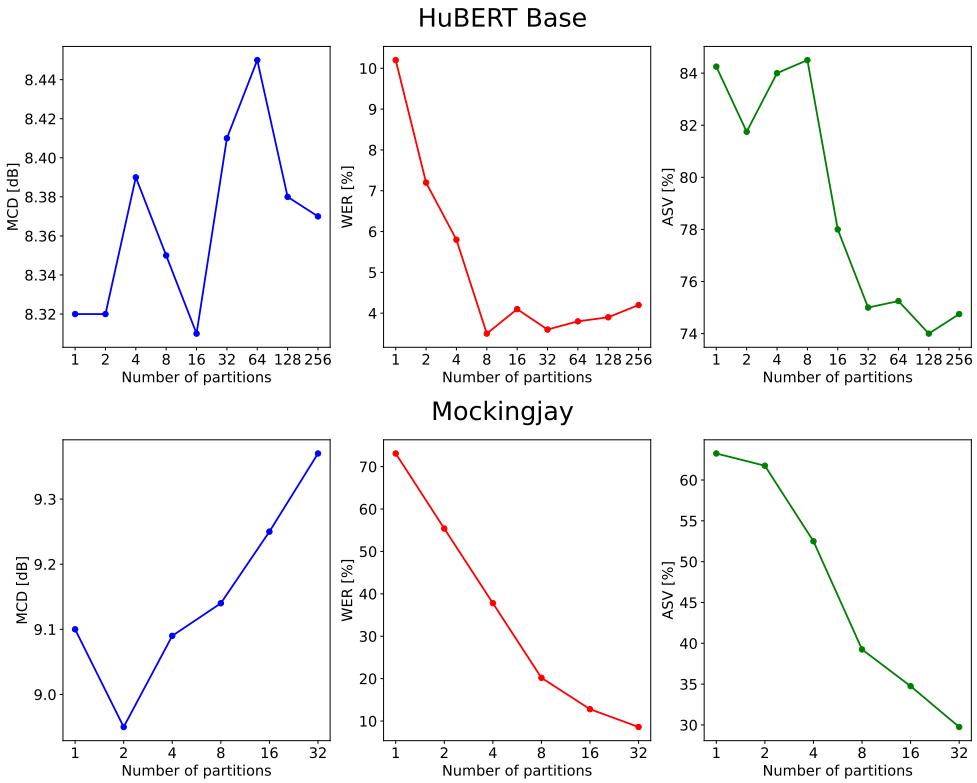


Figure 4.3: *Visualizing the effect of the number of partitions. Top: HuBERT Base. Bottom: Mockingjay.*

consistent across all numbers of clusters. However, using more k-means models in both cluster ensemble and PQ means loosening the speaker information bottleneck, which can harm the conversion similarity (ASV) as well as MCD. Finally, an interesting finding is that by only partitioning into two feature subvectors, the MCD and WER are still better than using an ensemble of three k-means models, suggesting that PQ is a more effective method than cluster ensemble. This is consistent with the finding in [39]. We thus use PQ in the following experiments.

Based on the observations in Table 4.8, the amount of speaker information leaked when the number of partitions increases is then studied. Table 4.9 shows the results

Table 4.9: *Results of HuBERT Base and Mockingjay varying the number of partitions (N_{PQ}) in the product quantization technique. The number of clusters is set to 200 in all k-means models. The task is any-to-any voice conversion, and the model is the Taco2-AR model.*

Upstream	N_{PQ}	MCD	WER	ASV
HuBERT Base	1	8.32	10.2	84.25
	2	8.32	7.2	81.75
	4	8.39	5.8	84.00
	8	8.35	**3.5**	**84.50**
	16	**8.31**	4.1	78.00
	32	8.41	3.6	75.00
	64	8.45	3.8	75.25
	128	8.38	3.9	74.00
	256	8.37	4.2	74.75
Mockingjay	1	9.10	73.1	**63.25**
	2	**8.95**	55.4	61.75
	4	9.09	37.8	52.50
	8	9.14	20.2	39.25
	16	9.25	12.8	34.75
	32	9.37	**8.6**	29.75

when varying the number of partitions using HuBERT and Mockingjay, and Figure 4.3 is a visualization of the overall trend.

For HuBERT Base, a diminishing returns effect in WER can be observed. That is to say, the WER stops to improve when N_{PQ} is large enough. The conversion accuracy

also stays at a similar level when N_{PQ} is small and starts to drop when N_{PQ} gets larger. These observations show that one can find an optimal N_{PQ} such that the WER is optimized while maintaining a similar level of conversion accuracy. However, for Mockingjay, both WER and ASV are monotonically decreasing, which means that such an optimal point cannot be found by only looking at these two metrics. As a result, MCD is used to find the optimal N_{PQ}.

Table 4.10: *Results on any-to-any voice conversion with continuous and discrete features over various upstreams. The results using continuous features are extracted from Table 4.3.*

Upstream	Continuous			Discrete		
	MCD	WER	ASV	MCD	WER	ASV
PASE+	9.85	4.2	8.00	8.92	81.7	74.00
APC	9.57	3.5	23.25	8.66	22.4	81.25
VQ-APC	9.43	4.0	22.00	8.42	21.0	85.50
NPC	9.39	4.4	21.00	8.78	46.0	74.50
Mockingjay	9.43	5.0	25.00	8.95	55.4	61.75
TERA	9.31	5.2	18.75	8.40	37.1	67.00
Modified CPC	9.61	4.1	10.75	8.69	13.8	75.50
DeCoAR 2.0	9.28	4.0	27.00	–	–	–†
wav2vec	8.77	3.5	40.00	8.34	15.2	**86.50**
vq-wav2vec	**8.47**	4.2	**73.25**	8.49	22.5	82.50
wav2vec 2.0 B.	9.03	3.2	27.00	8.90	54.3	75.75
wav2vec 2.0 L.	8.99	4.1	22.25	8.97	67.7	72.75
HuBERT B.	9.19	3.4	23.25	8.31	**4.1**	78.00
HuBERT L.	9.13	**3.0**	27.75	**8.23**	7.4	86.25

†: Fails to be trained.

4.4.8 Comparison of continuous and discrete features

Finally, we compare the results in the A2A setting when using continuous and discrete features. Since there are too many hyperparameters that can be tuned, the PQ technique is always applied and the number of clusters is set to 200. The best N_{PQ} between 1, 2, and 4 is searched using MCD.

Table 4.10 shows the results. It can be observed that the post-discretization process indeed serves as a strong speaker information bottleneck as the ASV scores of all S3Rs are significantly higher than the continuous counterpart. As described in Section 4.3.3, most S3Rs suffer from poor intelligibility even with the PQ technique. However, certain S3Rs still achieved an acceptable balance of intelligibility and conversion similarity, resulting in MCD values lower than that of the best performing continuous S3R (8.47 from vq-wav2vec), such as VQ-APC, wav2vec, HuBERT Base and HuBERT Large.

4.5 Discussion and conclusion

This chapter presented a comparative study of S3R-based VC. All experiments were based on S3PRL-VC, an extension of the S3PRL toolkit that focuses on the VC downstream task. The S3Rs were evaluated under the context of VC, and a series of in-depth analyses were carried out in various aspects including the synthesizer model type, different VC tasks, supervision, and discretization. The S3R-based systems were also compared with the state-of-the-art VC systems in VCC2020, and it was shown that there is still room for improvement in terms of quality and similarity.

Readers from different research communities can gain individual insights from this work. From the VC perspective, in S3PRL-VC, to meet the fast benchmarking requirement, some techniques that were shown to be effective were not applied, such as

fine-tuning target speaker-dependent vocoders [89, 132], training the synthesizer with waveform domain losses [81, 133], or fine-tuning the vocoder with ground truth aligned synthesis [42, 129, 134]. That is to say, the performance can be further optimized. In addition, applications to other VC tasks such as emotional VC, expressive VC, singing VC, and VC for speaking aid devices are also worth investigating.

From the S3R perspective, certain challenges are required by VC, such as the preservation of the spoken contents and the disentanglement of speaker information. It is therefore worthwhile to continue to use VC as a probing task when designing new S3R models.

Finally, it is worthwhile noting that VC has a special position in the context of the recent SUPERB [135] activities. SUPERB is a collection of benchmark resources that aims to evaluate S3Rs across various speech tasks, with an assumption in mind that different representations should outperform others in different tasks due to their pretext-task nature. However, in the original version which consisted of only 10 discriminative tasks, it turned out that wav2vec 2.0 and HuBERT outperformed all other S3Rs. This dominance was broken after the introduction of VC, where vq-wav2vec was shown to be the best in the A2O setting, due to its disentangling ability.

This finding has several important implications. First, it shows that VC can be used to examine the disentanglement performance of an S3R, and there is a need for disentanglement if one tries to develop a universal representation, which not yet exist. Also, it is expected that this work serves as a good initiative for future S3R researchers to emphasize the disentanglement performance of their model, without hurting the scores on other tasks like ASR and ASV. This could have a bigger impact on the community compared to pursuing incremental improvements on other tasks.

5 Ground-truth-free Application 1: Dysarthric Voice Conversion

In Chapter 3, a pre-training method for parallel voice conversion (VC) was proposed to tackle the problem described in Section 1.2.1, and in Chapter 4, self-supervised speech representation-based non-parallel VC was studied to address the problem described in Section 1.2.2. In the following two chapters, the focus is turned to the third problem as described in Section 1.2.3, which is to approach certain VC applications where the ground truth target for training is unavailable. In this chapter, a case study is conducted on the task of *dysarthric VC*.

5.1 Introduction

Dysarthria refers to a type of speech disorder caused by disruptions in the neuromotor interface such as cerebral palsy or amyotrophic lateral sclerosis [136]. Dysarthria patients lack normal control of the primary vocal articulators, resulting in abnormal and unintelligible speech with phoneme loss, unstable prosody, and imprecise articulation. Such a type of speech is referred to as **dysarthric speech**.

VC can be applied to improve the quality of life (QoL) of dysarthric patients by applying VC in two directions: dysarthric-to-normal (D2N) VC, and normal-to-dysarthric (N2D) VC. Figure 5.1 is an illustration. In the next two subsections (Sections 5.1.1

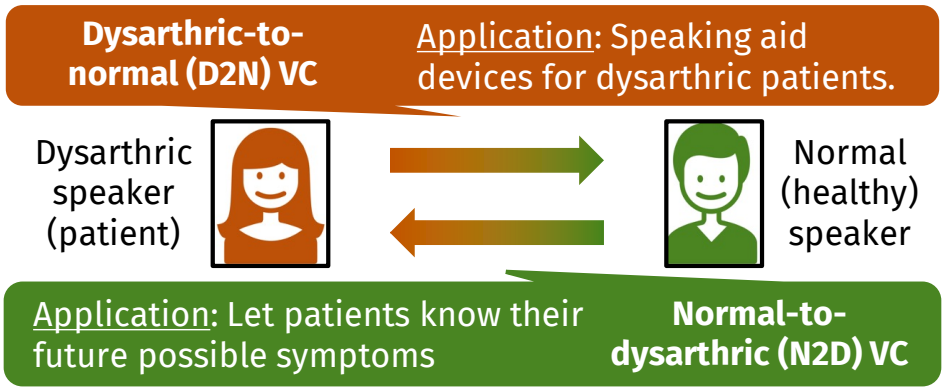


Figure 5.1: *Illustration of the dysarthric-to-normal and normal-to-dysarthric voice conversion.*

and 5.1.2), the importance of the respective directions as well as a brief literature review will be presented. An important property of these two tasks is **source speaker identity preservation**, and will be described in Section 5.1.3.

5.1.1 Dysarthric-to-normal voice conversion

D2N VC is needed because the ability of dysarthric patients to communicate with speech in everyday life is degraded. It is of urgent need to improve the intelligibility of the distorted dysarthric speech[1].

There have been several studies on D2N VC. Rule-based transformation based on signal processing [137] is limited in that each patient needs to be individually considered. Statistical approaches adopt models ranging from Gaussian mixture models [138], exemplar-based methods [139, 140] and deep neural networks [141–143]. However, as mentioned earlier, most of these methods did not take into consideration source speaker

[1]In the field of VC, orthogonal descriptions such as "naturalness" and "intelligibility" are often used, but the term "quality" is used interchangeably in this thesis.

identity preservation, which will be discussed in Section 5.1.3.

5.1.2 Normal-to-dysarthric voice conversion

On the other hand, N2D VC refers to the task of converting normal speech to dysarthric speech. A straightforward application is to improve automatic speech recognition (ASR) models against pathological speech by augmenting the training dataset with additional pathological data. Such augmentation could ease the low-resource constraints of a pathological ASR task. Previous works mainly employed frame-based VC models to convert the speech timbre, combined with extra procedures such as speed perturbation or dynamic time warping to modify the temporal structure and speaking rate [144,145]. These methods have shown promising improvements in ASR word error rates.

However, another important application that this chapter focuses on is informed decision-making related to the medical conditions at the root of speech pathology. For instance, an oral cancer surgery results in changes to a speaker's voice. The availability of a VC model that can generate how the voice could sound after surgery could help the patients and clinicians make informed decisions about the surgery and alleviate the stress of the patients.

So far, very few previous works have focused on VC for clinical usage. The first N2D VC system was presented in [146], which was a combination of a CycleGAN-based frame-wise VC model and a PSOLA-based speech rate modification process. However, this method suffers from audible vocoder artifacts brought by the extra PSOLA operation, and the inability to preserve the speaker identity of the control speaker. A different work [147] focused on dysarthric-to-dysarthric VC, by using a frame-wise VC model called HL-VQ-VAE [148]. However, the setup was not flexible in that (1) a

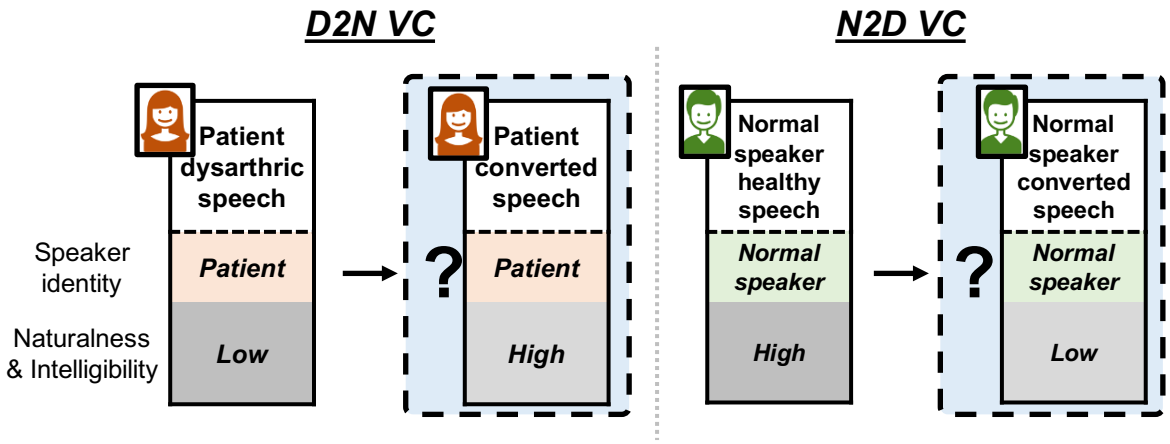


Figure 5.2: *Illustration of the common problem of the absence of ground truth training target in dysarthric-to-normal (left) and normal-to-dysarthric (right) voice conversion.*

severity-matched VC setup was required to avoid the need for varying speech rates, and (2) the method required a pathological source utterance, wherein real-world applications one might want to synthesize an arbitrary utterance from the normal source speaker.

5.1.3 Problem: absence of ground truth training target

Source speaker identity preservation is an important requirement for both D2N and N2D VC. Figure 5.2 provides an illustration. In D2N VC, given a speech utterance of the source patient with low naturalness and intelligibility, the goal is to synthesize the highly natural and intelligible counterpart while maintaining the speaker identity of the patient. Similarly, the goal of N2D VC is to simulate the dysarthric version of the source normal speaker's speech.

A common problem of these two directions is the **absence of ground truth training target**, illustrated as the blue squares in Figure 5.2. At a given time, collecting

a normal speech dataset of the patient is impossible. Or one should say, if such a collection process is possible, then this VC technique will not be needed anyway. The same logic applies to N2D VC.

As a start, one may try to apply the two representative VC techniques described in Chapter 2. However, neither the sequence-to-sequence (seq2seq) parallel VC approach nor the frame-based non-parallel VC method can solve the above-mentioned problem alone. For seq2seq VC models, a parallel training dataset is required, which is impossible to collect. On the other hand, frame-based VC will not be *sufficient* because of the frame-based conversion property. The inability of the frame-based VC framework to convert prosody is fatal because many unique attributes of dysarthric speech are related to prosody.

5.1.4 Solution: the cascade method

In this chapter, a solution is proposed to tackle the problem of the absent ground truth training target. The method is simply **cascading** a seq2seq model and a non-parallel frame-based model. It is referred to as the **cascade method** throughout this thesis. The main idea is to first use a seq2seq model to convert the source speech to the speech of a reference speaker. The intermediate speech is a by-product with the naturalness and intelligibility at a desired level, with an unwanted speaker identity of the reference speaker. Next, a frame-based, non-parallel VC model takes the intermediate speech with the identity of the reference speaker as input and restores the identity of the source speaker. An important assumption made here is that due to the frame-based constraint, the non-parallel VC model changes only time-invariant characteristics such as the speaker identity, while preserving time-variant characteristics, such as pronunciation. As a result, the converted speech has the speaker identity of

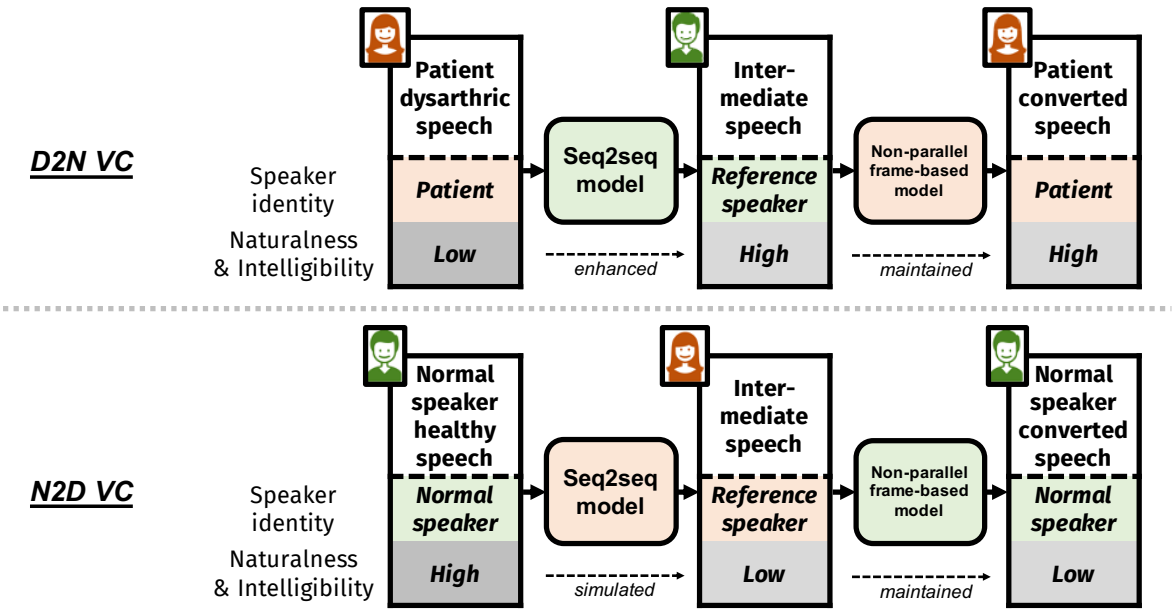


Figure 5.3: *Illustration of the conversion process in the proposed cascade method for preserving speaker identity. Top: dysarthric-to-normal voice conversion. Bottom: normal-to-dysarthric voice conversion.*

the source speaker while maintaining high intelligibility and naturalness.

In Section 5.2, the cascade method is described more in detail. Then, Sections 5.3 and 5.3 present the experimental evaluation of the cascade method on D2N and N2D VC, respectively.

5.2 The cascade method

5.2.1 General description

Suppose a training dataset of the source speaker is available. The first key ingredient to the proposed cascade method is the parallel counterpart, i.e. a training dataset from one or several reference speakers of identical contents. The second key ingredient, as

described in Section 5.1.4, is the assumption that due to the frame-based constraint, the non-parallel VC model changes only time-invariant characteristics such as the speaker identity while preserving time-variant characteristics.

Figure 5.3 is an illustration of the proposed cascade method. A seq2seq model first converts the input speech into that of a reference speaker, either to be more intelligible and natural or to simulate dysarthric speech. Then, a non-parallel frame-wise model restores the identity of the source speaker. The seq2seq model is based on the Voice Transformer Network (VTN) described in Section 2.1.2 with the text-to-speech (TTS) pre-training technique proposed in Chapter 3. The non-parallel frame-based model is based on the vector-quantized variational autoencoder (VQVAE)-based model proposed in *crank* [149], an open-source VC software that combines recent advances in autoencoder-based VC methods, including the use of hierarchical architectures, cyclic loss, and adversarial training.

5.2.2 One-to-many and many-to-one training of the sequence-to-sequence model

The original VTN was designed for one-to-one (O2O) VC, i.e., the model can only convert from one training source speaker to one training target speaker. That is to say, only the parallel utterance pairs of the source and the target speakers can be used. However, considering the difficulty of D2N VC and N2D VC, it would be beneficial to include more utterances to improve the performance. An advantage that can be taken on here is that data from more than one reference *normal* speaker can be used. While dysarthric patients can be rare, finding healthy speakers is relatively easy. In this subsection, based on such an advantage, two extensions of the original O2O VC are proposed for D2N VC and N2D VC.

In the case of D2N VC, the O2O VTN can be extended to a one-to-many version. Specifically, each parallel training utterance pair is composed of an utterance from the source patient and another utterance from a healthy speaker in the training speaker pool. Following the many-to-many VTN [150], a speaker embedding of the target speaker is concatenated to each of the hidden representation frames (as described in Section 2.1.1. The augmented feature sequence is then consumed by the decoder. The speaker embedding can be a simple one-hot embedding, while here the x-vector [151] is used.

It is worth investigating the choice of the reference speaker. Although the complete training dataset is parallel among the patient and all reference speakers, due to the difference in characteristics such as the speaking rate and F0 pattern, some speakers can be easier to convert to, compared to others. One may hypothesize that choosing a reference speaker with similar characteristics to the patient might make conversion easier. In the experimental evaluation section, an ablation analysis of how the choice of reference speaker affects the conversion performance in various aspects will be presented.

As for N2D VC, the model is trained in a many-to-one (M2O) fashion. While there is still only one source healthy speaker of interest, one may utilize data from multiple healthy speakers to improve the performance. Given a training utterance from any of the normal speakers, the VTN model is trained to convert to the predefined target dysarthric speaker. M2O training was also used in [152], except they used an auxiliary phoneme recognition regularization loss.

5.3 Experiments on dysarthric-to-normal voice conversion

5.3.1 Experimental settings

Datasets and implementation

The dysarthric dataset used in this experiment was a Mandarin dysarthric speech corpus provided by the Chi-Mei Hospital in Taiwan. It consists of dysarthric speech utterances read by a female patient. The prompts were from the TMHINT dataset [153], which contained 320 utterances each with ten Mandarin characters. The TMHINT dataset was designed to be phonetically balanced. For the reference speakers, the recordings of 17 speakers (13 male and 4 female speakers)[2] in the TMSV dataset [154] were used, where all speakers also uttered the TMHINT prompts. A 240/40/40 train/validation/test split was used, and all speech utterances were downsampled to 16 kHz. 80-dimensional mel-spectrograms with a 16 ms frame shift were extracted as the acoustic feature.

The implementation of the VTN model was the same as that described in Chapter 3. The TTS pretraining was conducted with the Sinica COSPRO multi-speaker Mandarin dataset [155], which is 44 hr long. The non-parallel frame-based model was based on a vector-quantized variational autoencoder (VQVAE)-based model implemented in *crank*, which can be accessed freely[3]. For simplicity, it is referred to as the *VAE* model in the rest of this chapter. Sinica COSPRO was used along with the TMSV and the patient's voice as training data for the VAE training. The Parallel WaveGAN (PWG) was used to generate from the converted mel-spectrogram the final waveform. The

[2]Speaker SP11 was excluded due to labeling error.

[3]`https://github.com/k2kobayashi/crank`

training data of PWG contained the recordings of the 18 TMSV speakers.

Evaluation metrics and protocols

For objective evaluation, the MCD described in Section 2.3 and the syllable error rate (SER) were used. Since calculating MCD requires the ground truth data, MCD was only calculated to evaluate the VTN model. On the other hand, to calculate the SER, a Transformer-based ASR model trained on the AISHELL-1 dataset [156] was first used to transcribe the converted utterance. Then, hey were then converted the characters into pinyin, and the tone was discarded to obtain the SER of the converted speech.

For subjective evaluation, the naturalness and conversion similarity described in Section 2.3 were evaluated. 11 native Mandarin speakers were recruited, and audio samples are available online[4].

5.3.2 Investigation of the choice of reference speaker

In this subsection, the hypothesis on the importance of the choice of reference speaker, which was discussed in Section 5.2.2, is examined. An objective evaluation is first conducted. Since two types of objective metrics (MCD and SER) were adopted in this section, it is worthwhile examining which is a more proper selection criterion. The O2M VTN model was trained for 2000 epochs, and the best-performing models were chosen based on MCD.

Figure 5.4 shows the results. First, since the patient is a female speaker, the gender might be a bias in reference speaker selection. From the figure, it could be observed

[4] `https://bit.ly/3sHxaGY`

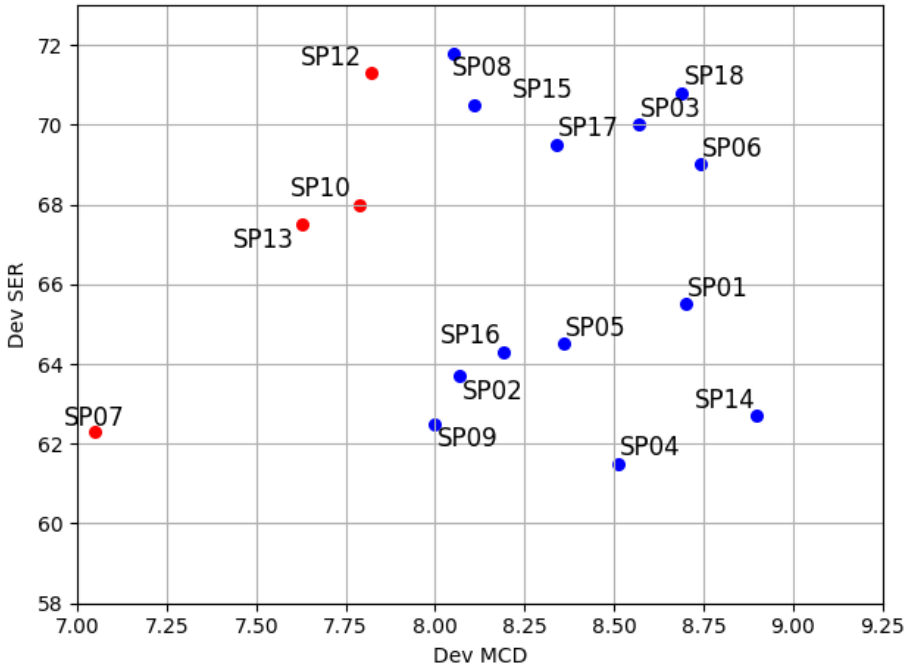


Figure 5.4: *Scatter plots of the MCD and SER scores of each speaker on the intermediate speech dev set in the dysarthric-to-normal voice conversion experiment. Both MCD and SER are the lower the better. Red and blue dots denote female and male speakers, respectively.*

that female reference speakers tend to yield lower MCD values. However, the SER scores did not differ much between genders, and none of the genders gave lower scores. It could be concluded that gender influences MCD but not SER.

Another observation is that the speaker with the lowest MCD score (SP07) did not necessarily give the lowest SER value (SP04 gave the lowest SER value), and vice versa. Thus, it is difficult to conclude which metric is better in the reference speaker selection process.

Table 5.1: *Results of **naturalness** evaluation in the dysarthric-to-normal experiments using the test set with 95% confidence intervals. Values are higher the better.*

Description	SP04	SP09	SP07	SP13
Patient	2.37 ± .19			
Reference speakers	4.99 ± .01			
Intermediate speech	3.29 ± .32	3.16 ± .27	3.45 ± .37	3.74 ± .27
Patient converted speech	2.42 ± .30	2.38 ± .41	2.65 ± .39	2.60 ± .35

Table 5.2: *Results of **similarity** evaluation in the dysarthric-to-normal experiments using the test set with 95% confidence intervals. Values are higher the better.*

Description	SP04	SP09	SP07	SP13
Patient	—			
Reference speakers	9% ± 7%			
Intermediate speech	8% ± 8%	8% ± 9%	30% ± 11%	25% ± 14%
Patient converted speech	45% ± 10%	45% ± 14%	49% ± 11%	42% ± 11%

5.3.3 Main results with subjective evaluation

In this subsection, the effectiveness of the proposed cascade method is verified with a subjective test. Since it is impractical to evaluate all converted samples of the 17 reference speakers, two speakers with the lowest MCD and SER values were chosen, i.e., MCD, SP07, and SP13 were chosen; for SER, SP04, and SP09 were chosen.

Table 5.1 shows the naturalness results. Although not significantly different, the reference speakers with lower MCD values (SP07, SP13) outperformed the other two speakers (SP04, SP09). Surprisingly, this trend holds for both intermediate and the

final converted speech. This result implies that listeners might have paid less attention to intelligibility, but valued other factors more. This also explains why the dysarthric speech, although with extremely low intelligibility, still yielded a MOS score of 2.37.

Table 5.2 shows the similarity results. The similarity scores of the final converted speech of the four speakers are not significantly different than each other, with SP07 slightly outperforming the other speakers.

Overall, the best-performing reference speaker was SP07, whose naturalness (2.65) and similarity (49%) scores were the best among all other speakers. One might also conclude that MCD seems to be a better metric for reference speaker selection, although more investigation needs to be made.

It is worthwhile discussing the ability of the proposed method to preserve the speaker identity of the source. Although the best similarity score was only 49%, feedback from the listeners suggested that it was easy to find the converted speech different from that of the dysarthric speech due to its special characteristics. Since the normal speech of the patient is impossible to obtain, it is essentially difficult to evaluate conversion similarity. Designing a better evaluation protocol is thus an important future work.

5.3.4 Degradation from the non-parallel frame-based VC model

As stated in Section 5.1.4, an important assumption in this chapter is that the pronunciation should be consistent throughout the non-parallel frame-based VC model. In this subsection, it is examined whether such an ability holds in the proposed method.

First, the preservation of intelligibility is examined by comparing the SERs of the original dysarthric voice to (1) the intermediate speech, which is the output of the VTN model, and (2) the final converted speech, which is the output of the VAE model.

Figure 5.5 shows the results. It can be observed that the assumption is only some-

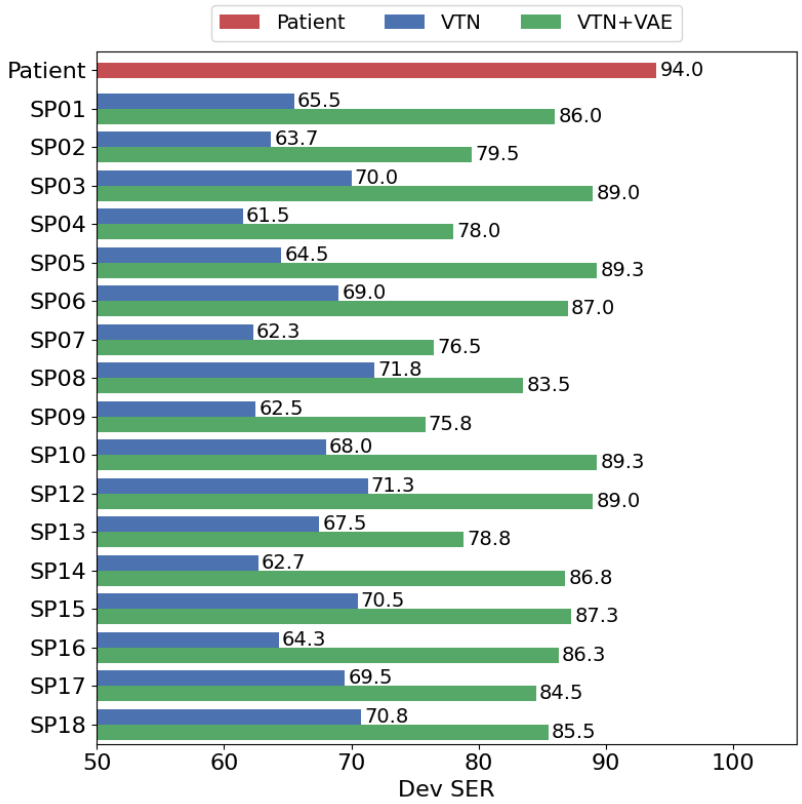


Figure 5.5: *SER values of the input dysarthric speech and the final converted speech of the dev set in the dysarthric-to-normal voice conversion experiment. The values are the smaller the better.*

what held, as all SER values of the final converted speech increased compared to those of the intermediate speech. An explanation is the insufficient disentanglement ability of the employed non-parallel frame-based model. As a result, a well-shared linguistic representation space between healthy and dysarthric speech cannot be learned.

Next, the naturalness consistency ability is examined. From Table 5.1, it could be observed that, regardless of which reference speaker, the naturalness scores degraded for almost 1 MOS point. This result again shows that the adopted non-parallel frame-based model could not guarantee such consistency.

Although the assumption is not completely satisfied in this experiment, the SER values from any reference speakers were all lower than that of the source dysarthric speech. Specifically, the final converted speech of speaker SP09 gave the lowest SER of 75.8, which was 18.2 points lower than the original 94.0. Also, as stated in Section 5.3.3, all reference speakers received a naturalness score higher than tof the patient. These results demonstrate the effectiveness of the cascade method.

5.4 Experiments on normal-to-dysarthric voice conversion

In D2N VC, the goal is rather clear: transform the source dysarthric speech into normal speech, such that it sounds as natural and intelligible as healthy speakers. On the other hand, as stated in Section 5.1.2, one important application of N2D VC is to let a patient know possible future symptoms. Dysarthria is a progressive process, which means that the **severity** level can gradually increase. Thus, one important goal of N2D VC is to **simulate** the severity level. In our experiment, the severity level is simply determined by the reference speaker, i.e., the dysarthric speaker. That is to say, given an input normal speech utterance and a reference speaker (represented by his/her training utterances), the converted speech generated by the N2D VC model is expected to have the same severity as the reference.

5.4.1 Experimental settings

Datasets and implementation

In this experiment, the UASpeech dataset [157] was used. UASpeech contains parallel word recordings of 15 dysarthric speakers and 13 normal control speakers. The training and test set consists of 510 and 255 utterances, respectively. Each dysarthric speaker is categorized into one of three intelligibility groups: low, mid, and high, which correspond to $0-25\%$, $25-75\%$, and $75-100\%$ subjective human transcription error rate (STER). The intelligibility of each speaker was judged by 5 non-expert American English native speakers. Two dysarthric speakers from each intelligibility group (high: M08, M10; mid: M05, M11; low: M04, M12) were chosen as the reference speakers for VC.

For each dysarthric speaker, a separate M2O VTN was trained using the data of that speaker and all 14 control speakers. Since UASpeech is an English dataset, TTS pretraining was conducted with M-AILABS judy [99], as in Chapter 3. The non-parallel frame-based VC method is also based on the VAE model proposed in crank [149], following the experiment in Section 5.3. In the preliminary experiments, it was found that training with only normal speech is better than training with a mix of dysarthric and normal speech. Thus, only utterances from the 13 control speakers were used to train the VAE model. Again, following Section 5.3, the PWG was used to generate from the converted mel-spectrogram the final waveform. The training data of PWG contained speech from all control speakers in UASpeech.

Evaluation metrics and protocols

For objective evaluation, the phoneme error rate (PER) was calculated with an HMM-based ASR model pre-trained on the TIMIT dataset [123], following [158]. First, the ASR model outputs phonemes of the converted speech. Then, a grapheme-to-phoneme tool[5] was used to transform the ground truth text into phonemes. Finally, the PER is calculated between the two phoneme sequences. In addition to the raw scores, the correlation coefficient r of the scores of the converted speech and the STER of the reference speech is also reported. For PER, the lower the better. For the correlation coefficient r, the higher the better.

For the subjective evaluation test, due to budget constraints, not all conversion pairs were evaluated, and audio samples can be found online[6]. The following dimensions were evaluated:

- **Naturalness**: The MOS test described in Section 2.3 was adopted, with the resolution modified from 1 to 0.5. At the beginning of the test, listeners were guided with an explanation of the definition of naturalness, followed by an example of normal and pathological (low severity) speech. Listeners were instructed to rate these both as 5 (highly natural). Each test asked the listener to rate 13 utterances for both pathological speakers of each severity (low, high, mid), leading to a total of 78 utterances. Subsequently, the experiment was repeated with the ground truth samples. In total, 30 native American English listeners were recruited.

- **Similarity**: The similarity test protocol described in Section 2.3 was adopted. Here, two scores are reported, which are the similarity scores to the source nor-

[5] An open-sourced tool was used: `https://github.com/Kyubyong/g2p`.

[6] `https://unilight.github.io/Publication-Demos/publications/n2d-vc`

mal speaker and reference dysarthric patient, respectively. Three pathological speakers (M04, M11, M10) which were shown to have recognizable characteristics in a previous study [147] were chosen. Furthermore, speech samples from two randomly sampled control speakers were also chosen. In total, 5 native American English listeners were recruited.

- **Severity**: Accessing the severity of dysarthric is a highly professional task. In the study, 3 trained speech-language pathologists (SLPs) were recruited. An AB evaluation study is carried out in this section. The listener is presented with two different synthesized utterances from two unknown speakers with different speech severity. The listener is asked to select the speech sample that is perceived as being more pathological. Since the desired severity is already known, the "accuracy" can be calculated, i.e., how often the SLP can correctly choose the sample that is supposed to be more severe. Due to the high cost of recruiting SLPs, only four conversion pairs were evaluated. Table 5.6 shows the four conversion pairs. For each pair, 20 utterances were rated.

5.4.2 Evaluation results

Objective evaluations

The PER results are shown in Table 5.3. It could be observed that the r value decreases dramatically from 0.83 to 0.68 after the non-parallel frame-based model. This finding is similar to that in Section 5.3.4, suggesting that the intelligibility preservation ability of the current VAE model is not sufficient.

Table 5.3: *PER and STER scores as well as the correlation coefficients r in the normal-to-dysarthric voice conversion experiment. The PER values are the lower the better, and the correlation coefficients are the higher the better.*

Severity group		**High**		**Mid**		**Low**		r
Reference speaker		M08	M10	M05	M11	M04	M12	
PER	Intermediate speech	58.7	55.1	84.1	71.8	79.6	103.4	0.83
	Converted speech	62.9	59.3	106.3	76.2	81.2	120.0	0.68
STER		7.0	7.0	42.0	38.0	98.0	92.6	1.0

Subjective evaluations

Table 5.4 shows the naturalness results. It could be observed that for natural speech, the MOS decreases with the increase in the severity level. This finding is similar to a previous study [147]. On the other hand, the converted samples are consistently rated as less natural than the natural ones (with $p < 0.001$ calculated with a Wilcoxon signed-rank test). However, it is worthwhile noting that, though not directly comparable to the results in [147], the MOS values are overall higher. This could be attributed to the use of the powerful seq2seq model.

Table 5.5 shows the similarity evaluation results. First, all scores in the "similarity to reference patient" column are less than 50%, suggesting that there is no remaining reference patient speaker identity in the final converted speech. However, at the same time, all scores in the "similarity to source normal speaker" column are also less than 50%, except for the "M10 $\rightarrow$ CF03" pair. Before concluding that the identity preser-

Table 5.4: *Naturalness results with 95% confidence intervals in the normal-to-dysarthric voice conversion experiment. All values are the higher the better.*

Severity level	**Control speakers**	**High**	**Mid**	**Low**
Natural speech	3.93 ± 0.54	3.92 ± 0.54	2.86 ± 0.89	2.32 ± 1.16
Converted speech	-	2.70 ± 0.95	2.28 ± 1.03	1.94 ± 1.21

vation requirement is not satisfied, it is worthwhile noting that, from the feedback of listeners, it is again difficult to evaluate similarity by listening to a dysarthric speech sample and a normal speech sample. This problem is also present in Section 5.3.3.

Finally, Table 5.6 shows the severity results. First, the SLPs always perceived the more severe speakers as more severe. This is evident by the fact that each entry in Table 5.6 is over 50%. Compared to the natural speech, the accuracy of the intermediate speech is very similar, indicating that the intermediate speech simulates the severity aspect well. However, a large degradation can be observed in all conversion pairs in the converted speech stage, suggesting that the distortion caused by the VAE model makes it harder for SLPs to correctly distinguish the severity. This conclusion is again consistent with that in Section 5.3.3.

5.5 Conclusions and Discussions

In this chapter, the application of VC to dysarthric voice conversion was studied. To solve the problem of the absence of ground truth training data, a cascade method that combines (1) seq2seq modeling based on parallel data and (2) non-parallel frame-based modeling realized with a VAE model was proposed.

In the D2N VC experiments, it was demonstrated that the cascade method could

Table 5.5: *Similarity results with 95% confidence intervals in the normal-to-dysarthric voice conversion experiments. For similarity to source normal speaker, the higher the better; for similarity to reference patient, the lower the better.*

	Similarity to source normal speaker	Similarity to reference patient
M04 → CM05	20% ± 10%	32% ± 12%
M11 → CM09	37% ± 13%	43% ± 13%
M10 → CF03	55% ± 13%	8% ± 7%
M04 → CM04	33% ± 12%	27% ± 12%
M11 → CM10	23% ± 11%	32% ± 12%
M10 → CF02	48% ± 13%	10% ± 8%
Ideal	100%	0%

improve the quality and preserve the source speaker identity to a certain extent. Yet, the VAE model does not guarantee quality consistency, leading to sub-optimal quality and speaker identity preservation ability.

In the N2D VC experiments, it was found that (1) a better naturalness was achieved compared to that in the previous work [147], (2) the cascade method was able to mimic the severity characteristics linearly according to the SLPs, and (3) the quality consistency and speaker identity preservation ability can still be improved.

To conclude this chapter, the following is a list of three possible future directions.

- **Improving the seq2seq model.** Compared to the dataset used in Chapter 3, dysarthric speech contains more variation and is thus harder to model for a complicated model like the seq2seq network. From in-lab studies, it was found that the speech samples generated by the seq2seq model suffer from problems like

Table 5.6: *Severity results with significance levels calculated using a binomial test in the normal-to-dysarthric voice conversion experiments. The severity of each speaker is specified in the parentheses. ***: $p < 0.001$; *: $p < 0.05$.*

Speaker pair	Natural speech	Intermediate speech	Converted speech
M04 (low) vs M05 (mid)	95% ***	85% ***	53%
M05 (mid) vs M08 (high)	90% ***	95% ***	80% ***
M12 (low) vs M11 (mid)	93% ***	85% ***	75% *
M11 (mid) vs M10 (high)	98% ***	95% ***	68% *

missing and repeating words, which are common problems of a not well-trained seq2seq model. Possible improving techniques include text supervision [159], data augmentation [152], or non-autoregressive modeling [160].

- **A non-parallel frame-based model with better quality preservation ability.** The employed VAE model was shown to be insufficient in preserving intelligibility and naturalness. It is worthwhile resorting to other frame-based models such as the recognition-synthesis-based methods described in Section 2.4.

- **Better evaluation protocols.** From Tables 5.1, 5.2, 5.4 and 5.2, it could be observed that the confidence intervals were all much larger than those seen in Chapters 3 and 4, indicating that there were large disagreements between the listeners. Even the SLPs, who underwent professional training, provided feedback that the tests were quite difficult. It is worthwhile to re-examine the evaluation protocol designs.

6 Ground-truth-free Application 2: Foreign Accent Conversion

In this chapter, continuing the study of the third problem stated in Section 1.2.3, the task of foreign accent conversion (FAC) is studied[1].

6.1 Introduction

The FAC task in this chapter is to, given an accented speech utterance spoken by a non-native source speaker, generate a native-sounding version with the same speaker identity as the source speaker. Applications of FAC include computer-aided language learning [5, 161, 162] and entertainment such as movie dubbing [163]. As stated in Section 1.2.3, FAC suffers from the same ground truth training data absence problem, as it is impossible to collect native speech from a non-native speaker.

In the FAC literature, most works tried to utilize accent-independent features to decompose accents from voice identity. For instance, early attempts made use of articulatory trajectories (e.g., lips and tongue movements) [164–166] and vocal tract length normalization [167]. More recently, more simplified features such as phonetic posteriorgrams (PPGs) [168, 169] and text [170] are combined with advanced deep neural

[1]Readers should note that the term "accent conversion" can be referred to many different tasks in the literature. While many have used this term to refer to the conversion between different accents or from native to accented speech, in this chapter the focus is on the task of "de-accenting".

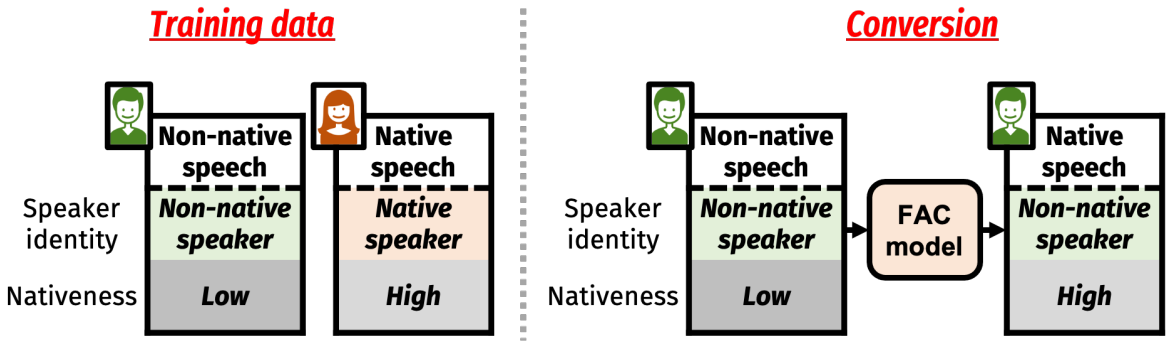


Figure 6.1: *Left: the training data, which is a parallel corpus between the source non-native speaker and a reference native speaker. Right: the goal in the conversion phase of FAC. The nativeness is expected to be increased while maintaining the speaker identity.*

network architectures, especially sequence-to-sequence (seq2seq) VC models, whose ability to model segmental and prosody features simultaneously play a crucial role in FAC.

However, only very few works have tried to address ground-truth-free FAC [171,172][2]. In these works, the main idea is similar to that stated in Section 5.2.1. The first ingredient, as depicted in the left-hand side of Figure 6.1, is to collect a training corpus from the source non-native speaker and then collect the native counterpart from a native reference speaker with the same prompt set. Then, a combination of state-of-the-art VC methods for disentangling the speaker and content is designed to achieve FAC. In addition to the cascade method described in Chapter 5, some previous works in the field of FAC also share a similar idea [171, 172]. However, these works were developed in parallel, and comparisons were often conducted on a system-to-system basis, leading to a lack of comparison and understanding of each method.

[2]Some previous works [171] used the term 'reference-free', but it might be confusing since a reference speaker is needed for training. Therefore, the term "ground-truth-free" is used in the rest of the paper, where "ground-truth" refers to the ground-truth used as the training target.

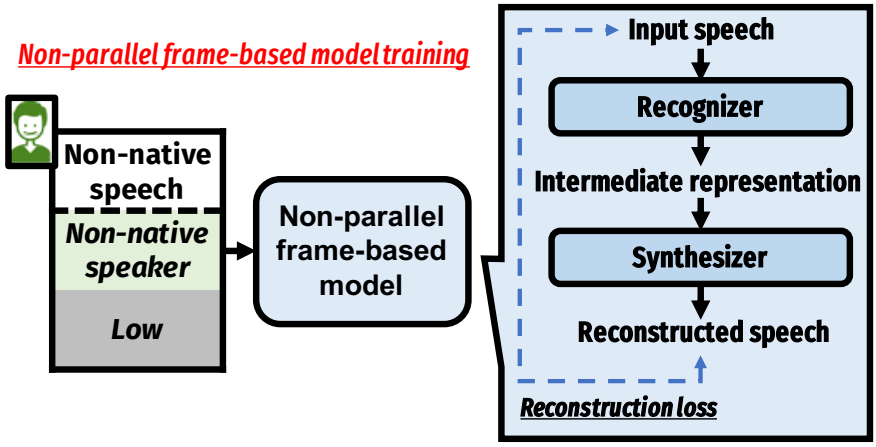


Figure 6.2: *The non-parallel frame-based voice conversion model and its training process used in the foreign accent conversion experiment.*

In this chapter, the aim is to systematically evaluate three ground-truth-free methods for FAC. Experiments are conducted in a unified setting using a shared database, model architecture, and waveform synthesizer. A subjective evaluation test is conducted to assess three different aspects of the synthesized samples, namely naturalness, speaker similarity, and accentedness. As will be shown in the experimental evaluation section, it was found **no single method was significantly better than the others in all evaluation axes**, which is in contrast to conclusions drawn in previous studies [172]. Results of an objective intelligibility measure which was used in previous studies [171] showed that it might not correlate well to subjective accentedness. Finally, to promote reproducible FAC research, the implementation is open-sourced to help future researchers improve upon the evaluated systems[3].

[3] `https://github.com/unilight/seq2seq-vc`

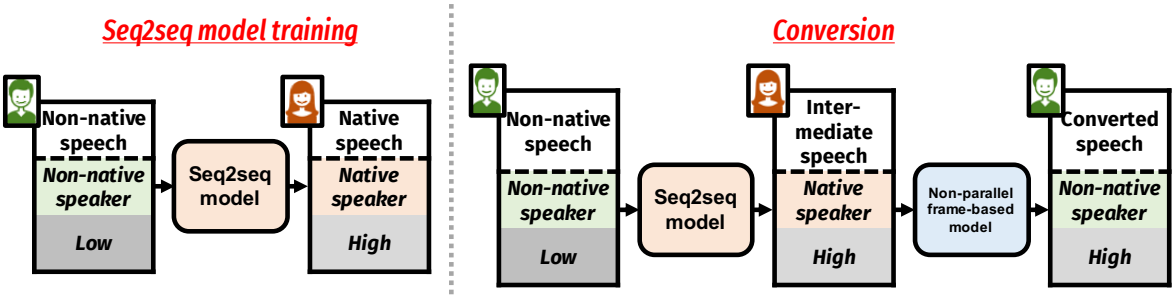


Figure 6.3: *Illustration of the training and conversion processes of the cascade method for foreign accent conversion.*

6.2 Evaluated methods

The three methods to be evaluated share two common materials. The first material is a parallel dataset between the non-native speaker and a reference native speaker, as described in Section 6.1. The second material is a non-parallel frame-based VC model. Different from the autoencoder-based method used in Chapter 5, to reflect the quality degradation problem in Section 5.5, the recognition-synthesis (rec-syn)-based method described in Section 2.2.2 is used. The training of the rec-syn-based VC model requires only the data of the speaker to be synthesized. Thus, only the dataset of the source non-native speaker is used, as illustrated in Figure 6.2. This model will then be fixed in the subsequent training processes of all three methods. In the following subsections, the detailed procedures of all three methods will be described.

6.2.1 Method 1: cascade

The cascade method is essentially identical to that described in Section 5.2. Although it was applied to dysarthric VC, since dysarthric VC and FAC share a common problem of lacking ground-truth training target, here it is examined whether it could be applied

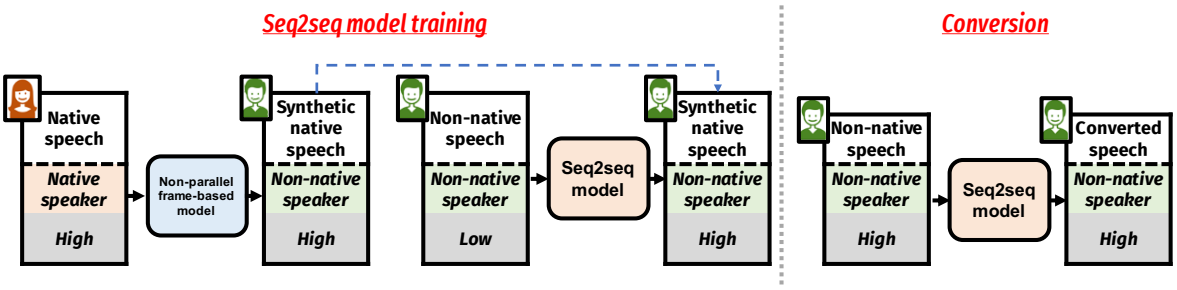


Figure 6.4: *Illustration of the training and conversion processes of the synthetic target generation method for foreign accent conversion.*

to FAC.

In the cascade method, a seq2seq model is trained to map from the source non-native speech to that of the reference native speaker. During conversion, the source speech is first sent into the seq2seq model to get the first stage converted speech. Although the nativeness is improved, the speaker's identity is changed into that of the reference speaker. Therefore, the non-parallel VC model is then used to change the identity back to that of the native speaker, while maintaining the pronunciation.

6.2.2 Method 2: synthetic target generation (STG)

The second method is called synthetic target generation (STG) [171]. The main idea is to use the non-parallel frame-based VC model to convert the training dataset of the native speaker. The resulting *synthetic training target* has (1) the speaker identity of the non-native speaker, and (2) the nativeness of the reference native speaker, depending on the pronunciation preservation ability of the non-parallel frame-based VC model. This step gives the name "*synthetic target generation*". Then, the seq2seq model is simply trained using the non-native training set as the source, and the synthetic native speech with the speaker identity of the same non-native speaker as the target.

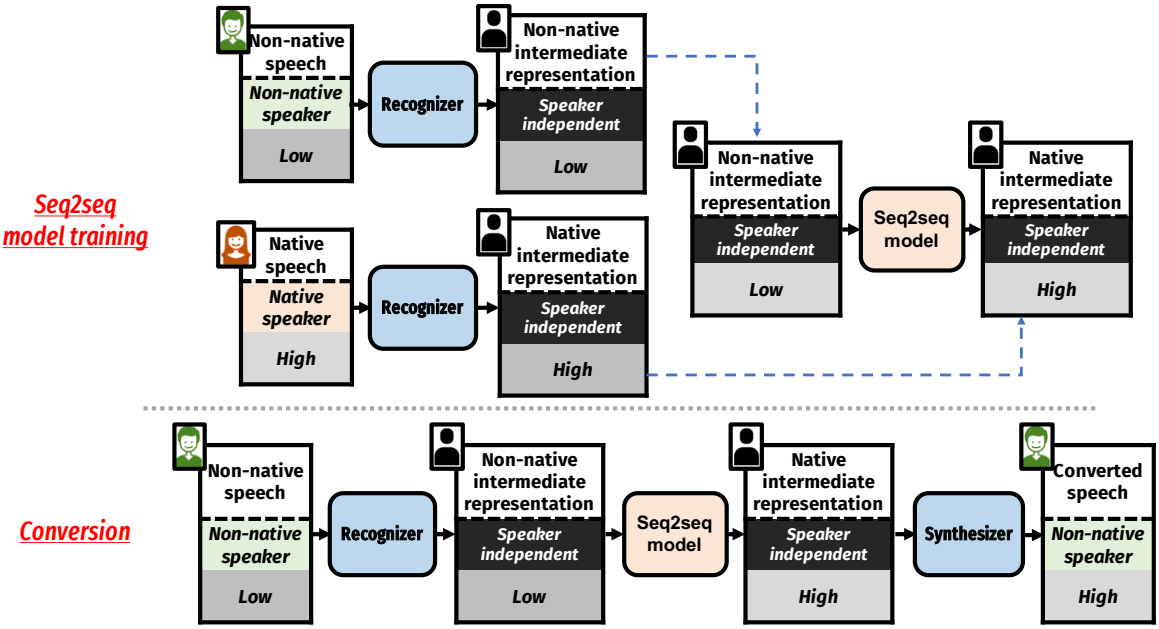


Figure 6.5: *Illustration of the training and conversion processes of the latent space conversion method for foreign accent conversion.*

During conversion, different from the cascade method where two models are needed, only the seq2seq model is used to generate the de-accented speech with the identity of the non-native speaker.

6.2.3 Method 3: latent space conversion (LSC)

The third method is called latent space conversion (LSC) [172][4]. The main idea is to utilize the speaker-independent attribute of the intermediate representation in rec-syn-based VC models. First, the recognizer transfers the training speech utterances of the source non-native and target native speakers from the speech space to the intermediate representation (or latent feature) space. Then, the seq2seq model is trained to map

[4]The term "latent feature" and "latent space" is used interchangeably with "intermediate representation" and "intermediate representation space", respectively.

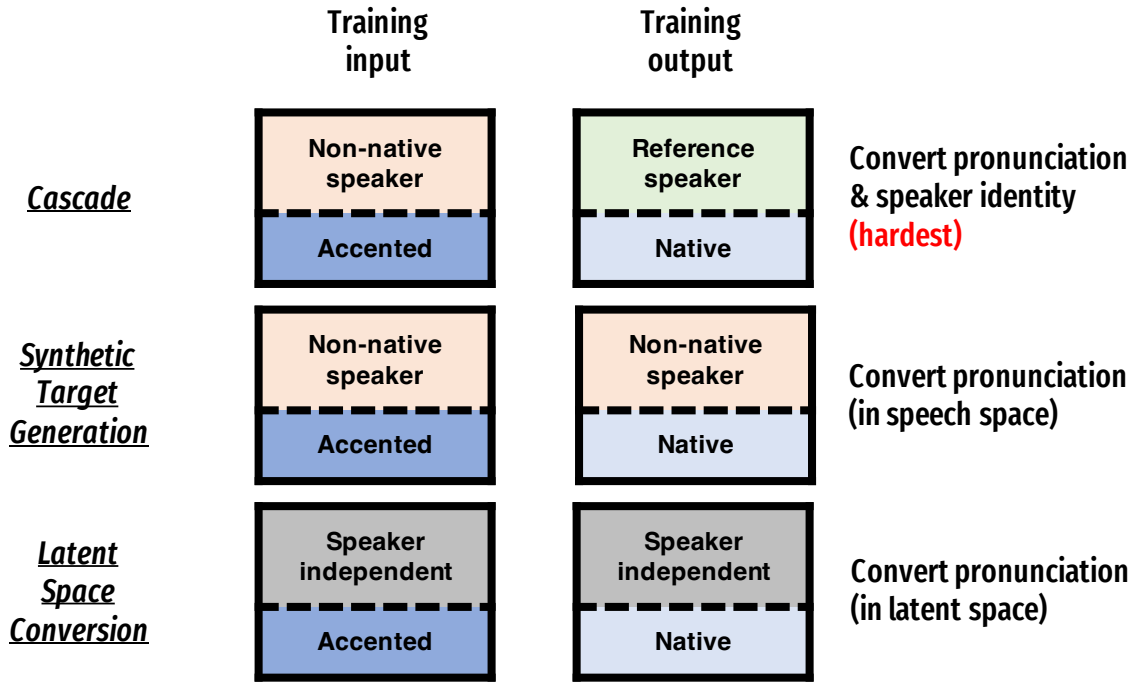


Figure 6.6: *Illustration of the difference in the training of the sequence-to-sequence model of the three evaluated methods in the foreign accent conversion experiment.*

the source latent features to the target latent features.

During conversion, the intermediate representation of the source non-native speech is first extracted and transformed to the native counterpart using the seq2seq model. Finally, the synthesizer of the rec-syn-based VC model is used to inject the identity of the non-native speaker into the converted latent features to generate the final converted speech.

6.2.4 Difference between the three methods

As these methods might be seemingly complicated in their ways, the difference lies in the **training input and output of the seq2seq model**. Figure 6.6 depicts such

a difference. First, in the cascade method, the seq2seq model needs to simultaneously convert the speaker identity and the nativeness, which is considered to be the most difficult. In contrast, STG first changes the speaker identity of the native training set, such that the seq2seq model only needs to model the pronunciation pattern. Finally, to further ease the job of the seq2seq model, LSC projects the training datasets onto the latent space, which is speaker-independent. Based on this observation, although it is difficult to determine whether the learning of the seq2seq model in STG is harder than in LSC, it could at least be hypothesized that the performance of cascade is worse than those of STG and LSC. This hypothesis will be examined in Section 6.3.3.

These three methods also have their weakness. For instance, during the conversion phase, not only cascade but also LSC does the input speech pass through a pipeline consisting of multiple modules, suffering from potential error propagation. STG, on the other hand, requires only the seq2seq module during conversion and thus does not suffer from error propagation, but the synthetic target data inevitably contains artifacts. The performance of the seq2seq model is then bounded by how imperfect the synthetic data is. With these unique limitations, readers should note that it is difficult to compare these methods.

6.3 Experimental evaluation results

6.3.1 Experimental setting

The datasets used in this experiment are L2-ARCTIC [173] and ARCTIC. Specifically, the non-native and reference native speakers are THXC (Chinese male) and bdl (English male) from the L2-ARCTIC [173] and ARCTIC datasets, respectively. There are 1032/50/50 training/development/testing parallel utterances respectively, and the

total duration of the training set is around one hour. All samples are in 16 kHz. The downsampled LJSpeech dataset [10] was used for the seq2seq model pertaining.

The seq2seq model was again based on the VTN model described in Chapter 3 with the same TTS pre-training dataset, M-AILABS judy. For the rec-syn-based model, the Taco2-AR model described in Chapter 4 was adopted. As an ablation study, two types of intermediate representations were investigated: (1) PPG extracted from a Conformer automatic speech recognition (ASR) model trained on the LibriSpeech dataset (960 hours), and (2) vq-wav2vec [116], a self-supervised speech representation whose extractor was trained on the LibriLight [118] dataset (60k hours). The Parallel WaveGAN [104] was used to generate from the converted mel-spectrogram the final waveform. It was trained with only the training set of the source non-native speaker. As the implementation is open-sourced[5], interested readers are encouraged to refer to the source code for detailed hyperparameters.

Following previous works [171, 172], a subjective evaluation test was conducted on three axes. Audio samples are available online[6].

- **Naturalness** was evaluated using a MOS test, as described in Section 2.3. Samples of the source non-native and reference native speech were also included. The Amazon Mechanical Turk (MTurk) crowd-sourcing platform was used to recruit 65 workers, and each of them rated 20 samples.

- **Similarity** was evaluated in the same fashion as described in Section 2.3. The same 65 workers from Amazon MTurk who participated the naturalness test were also asked to rate the similarity.

- **Accentedness** was also evaluated using a MOS test, with the only difference in

[5] See footnote 3

[6] `https://unilight.github.io/Publication-Demos/publications/fac-evaluate`

Table 6.1: *Objective and subjective evaluation results with 95% confidence interval of samples from the evaluated methods, source and target.*

Method	Extractor	CER/WER	Naturalness ↑ (1-5)	Similarity ↑ (0% -100%)	Accentedness ↓ (1-9)
Source (non-native)		5.3/12.3	4.18±0.19	–	6.06±0.38
Cascade	vq-wav2vec	29.1/52.5	3.17±0.23	28.7%±6.7%	5.41±0.32
	PPG	30.4/52.7	3.50±0.22	45.7%±7.3%	4.18±0.30
STG	vq-wav2vec	25.3/45.0	3.23±0.21	37.0%±7.0%	5.27±0.31
	PPG	17.7/40.9	3.66±0.20	57.3%±7.8%	4.36±0.32
LSC	vq-wav2vec	33.4/52.5	3.65±0.25	36.0%±7.0%	4.61±0.32
	PPG	9.8/19.5	3.64±0.22	43.8%±7.5%	3.95±0.31
Target (native)		1.3/4.3	4.42±0.18	–	1.49±0.21

that following [171], a 9-point scale was used. In addition, due to the difficulty of the task itself, instead of using Amazon MTurk, an in-lab study was conducted by recruiting 19 listeners to each listen to 40 samples.

Only one objective measure was calculated in this experiment: the character/word error rate (CER/WER) obtained by running an ASR model on the speech samples. The same pretrained ASR model based on wav2vec 2.0 was used, following Chapter 4.

6.3.2 Design choice of the non-parallel frame-based model

The ablation study on the effectiveness of PPGs and vq-wav2vec as the intermediate representation in the task of FAC is first presented. The results are shown in Table 6.1.

In all three subjective evaluation axes (naturalness, similarity, and accentedness) and all three evaluated methods, using PPG was almost always significantly better than using vq-wav2vec. The only exception was that the naturalness scores were nearly identical when using vq-wav2vec and PPG in the LSC scenario.

In Chapter 4, it was already shown that PPG trained on large-scale datasets could outperform vq-wav2vec in terms of naturalness and similarity. From Table 6.1, the superiority of PPG in accentedness further verifies the importance of linguistic supervision in the training of the recognizer. In the rest of the section, the focus will be on the results of the three methods using PPG.

6.3.3 Main results of the three evaluated methods

In this subsection, the results in Table 6.1 will be analyzed using the three subjective evaluation axes. Meanwhile, the performance tendency will be examined using the hypothesis described in Section 6.2.4.

In terms of naturalness, Table 6.1 shows that, as the confidence intervals overlap, there was no statistically significant difference between the three methods. This suggests that naturalness is not affected by the difficulty of the seq2seq mapping.

Next, for similarity, STG is significantly better than cascade and LSC. This result somehow verifies the above-mentioned hypothesis. Also, the relatively low similarity of LSC implies that the assumption of speaker independence of the latent features may be invalid in the context of FAC.

Finally, in terms of accentedness, the only significant difference that can be observed is the superiority of LSC over STG. Although this does not match the hypothesis, note that the accentedness score of LSC is significantly better than cascade and STG when using vq-wav2vec. This suggests that LSC is more robust to the choice of the

intermediate representation.

In general, no single method can significantly outperform the others in all three subjective evaluation axes. This is probably one of the most important messages in this study, as it contradicts the conclusion in [172].

6.3.4 Is character/word error rate a proper objective measure for FAC?

The feasibility of using objective measures to predict subjective results is a long-standing problem in VC research [94, 174]. Developing such a measure allows for inspecting the performance during system development without the expensive subjective evaluation process. Some previous works on FAC reported CER/WER as an indirect measure of accentedness, with the expectation that reducing accentedness can also reduce the error rates. The validation of such an expectation will be investigated in this subsection.

With the 8 data points in Table 6.1, the linear correlation coefficients between accentedness and CER/WER are 0.413 and 0.442, respectively. It can be then inferred that there is a weak yet insignificant correlation between accentedness and CER/WER. It is therefore concluded that there are other factors than intelligibility when it comes to accentedness, thus using CER/WER solely as an objective measure for FAC is unreliable.

6.4 Discussions and Conclusions

In this chapter, three methods for ground-truth-free FAC were evaluated systematically. Experiments were carried out in a unified setting, and subjective tests were

conducted in terms of naturalness, speaker similarity, and accentedness. In addition to the detailed discussion of each method and evaluation axis presented in Section 6.3.3, the most important message that the evaluation results show is that **no single method was significantly better than the other two in all evaluation axes**.

While this may arise from the insufficiency of the evaluated methods, it is also worthwhile questioning whether the evaluation protocols adopted in this chapter are proper. Although the subjective evaluation protocol was designed to be as close to that adopted in previous works [171,172] as possible, the large confidence intervals observed in Table 6.1 suggest that a more reliable protocol needs to be developed. Qualitatively, listener feedback suggests that a 9-point scale test as used in [171,172,175] is too fine-grained to give precise ratings. Also, while it is easy to tell whether a sample is native or not, rating the degree of accentedness is rather difficult. Comparative measurements such as preference tests might be more suitable, as advised in [176].

Also, listeners mentioned that even as native English speakers, it was difficult to rate accentedness. One way to improve this is to provide a training section containing utterances with different levels of accentedness, as the one provided in [177]. Another alternative is to directly recruit linguistics or educators as in Section 5.4.1, as someone with in-depth professional knowledge may make judgments more confidently.

7 Conclusions

7.1 Summary of This Thesis

This thesis revolved around the task of voice conversion (VC), and the problem of interest was to tackle the eternal scarcity of training data to learn a good mapping function. The main idea of this thesis was to apply pre-training, which tries to transfer knowledge from a machine learning model trained with a larger dataset in another domain. In this thesis, pre-training was applied to improve the performance and training data efficiency of VC, as well as the application to ground-truth-free VC problems.

In Chapter 3, the focus was on improving sequence-to-sequence (seq2seq) VC modeling, which was on the **performance** dimension. The proposed method was a pre-training technique based on two popular tasks in speech processing, namely text-to-speech (TTS) and automatic speech recognition (ASR). From an information perspective, these two tasks are by nature suitable sources of knowledge transfer, and the abundant resources (i.e., available training data) in these two tasks also motivate the use of these two tasks as the pretext task. The main contribution of this chapter was to boost the robustness of seq2seq VC models such that when trained with only five minutes of parallel training data, a naturalness mean opinion score (MOS) of 4.11 and a similarity score of 68% could be achieved with the proposed TTS pre-training technique.

In Chapter 4, the focus was on improving recognition-synthesis (rec-syn)-based non-

parallel VC, which was on the **training data efficiency** dimension. A large-scale study was carried out to examine the effectiveness of applying various self-supervised speech representations (S3Rs) as the intermediate representation in rec-syn-based VC. Self-supervised learning is attractive in unlocking the use of unlabeled datasets for pre-training, which has become a dominant training paradigm in many research fields in deep learning. The main contribution was the development of S3PRL-VC, a series of academic research activities including a toolkit, a benchmark, and a series of evaluation results of S3R-based VC. The unified task design, model architecture, and evaluation metrics are beneficial for future S3R researchers to conveniently test their newly developed S3R model. The large-scale study also brought various fruitful insights to not only the S3R but also the VC community. For the S3R community, by utilizing the characteristics of VC, the ability of each S3R model to disentangle speaker and content information was systematically investigated. For the VC community, it could be expected that the first unified comparative study of S3R-based VC could guide researchers toward better design choices when building their systems.

The third dimension was to solve a typical VC application category, where the ground truth training target is unavailable. In this category, instead of converting the speaker's identity as in most VC research, the goal is to transform a certain attribute while maintaining the identity. In these applications, it is impossible to collect the training target for the VC model training, so seq2seq models cannot be applied. Also, the attribute is often correlated to prosody, which cannot be well converted by frame-based VC models. In Chapter 5, as the first case study, dysarthric VC was investigated. As an initial investigation, a cascade method, which combines seq2seq and rec-syn-based VC modeling, was proposed. The main contribution of this chapter was to show initial investigation results. In dysarthric-to-normal (D2N) VC, a naturalness mean

opinion score (MOS) of 2.65 and a similarity score of 49% were achieved. In normal-to-dysarthric (N2D) VC, it was initially shown that the proposed method could convert the severity linearly, i.e., listeners could judge which converted speech sample was more severe than the other. Discussions also revealed the importance of reflecting on the evaluation experiment design.

In Chapter 6, the task of foreign accent conversion (FAC) was chosen as the second application study. In addition to the cascade method, two additional methods that were previously proposed in the FAC literature were taken into consideration, as they all combined seq2seq and rec-syn-based VC modeling. a systematical experimental evaluation was conducted to examine the naturalness, similarity, and accentedness aspects of the three methods. The main contribution of this chapter was to show that no single method was superior to the others, which contradicted the conclusion derived in previous studies. Similar to Chapter 5, the results again suggested the importance of a better evaluation protocol design.

7.2 Future Work

The results of the voice conversion challenge (VCC) 2020 [9] made some people think that VC is a "solved task", as the top system achieved a similarity score that was not significantly different from natural speech samples, and the naturalness also almost reached that level. In this thesis, the thesis scope illustrated in Figure 1.5 provided a new perspective on the current VC research trend. Based on the results and insights derived from this thesis, several potential research directions can be pursued in the future, as will be discussed in the following subsections.

7.2.1 Low-latency, real-time sequence-to-sequence VC

There are many applications where low-latency, real-time (LLRT) VC is demanded. A system is considered *real-time* if the real-time factor (RTF), which is the ratio between (1) the time taken to process the input and (2) the input duration, is smaller or equal to 1. On the other hand, latency refers to the time difference between the input and the output, and in speech communication, the threshold of a low-latency system is around 100 ms[1].

The D2N VC described in Chapter 5 is one typical application that requires the VC system to be LLRT. One could easily imagine that hand-held speaking aid devices that support D2N VC shall demand such a fast, instant processing speed. On the other hand, certain applications of the FAC task described in Chapter 6 also require the VC system to be LLRT. For instance, in international call centers that recruit non-native customer service agents, a FAC system that de-accents the talkers' speech needs to operate immediately with low latency.

There has been a line of work on developing LLRT VC systems [178–180]. However, these systems are mostly based on frame-based models. As discussed previously, the methods used in Chapters 5 and 6 adopted seq2seq models to model prosody, which plays an important role in applications like dysarthric VC and FAC. An initial attempt was made to develop LLRT seq2seq VC [181], which can be seen as an important related work to further apply to dysarthric VC and FAC.

[1]Another term *streaming* refers to the ability to process one frame in a time interval smaller than the frame shift. For example, if the frame shift is set to 10 ms, then the time to generate one frame of speech should be smaller than 10 ms.

7.2.2 Controllable intermediate representation for various VC applications

In Chapters 4 and 6, the phonetic posteriorgram (PPG) was shown to be superior to S3Rs in rec-syn-based VC, even for applications that are beyond speaker conversion. An explanation was that unnecessary information in the S3Rs was not well disentangled, giving the synthesizers a hard time to be well-trained. This result is somewhat counter-intuitive, as one might assume that the larger the pre-training dataset is, the better the performance is in the downstream task. Also, the "unnecessary" information in the S3R could be expected to be beneficial in certain applications. For instance, in the singing voice conversion challenge 2023 (SVCC2023) [182], several top-performing systems adopted advanced S3Rs instead of PPGs. It is therefore of significant interest to design a controllable intermediate representation, whether based on self-supervised learning or not, such that one could decide what information to be kept in the representation, depending on the target application.

7.2.3 Better evaluation design for VC

In almost all subjective evaluation tests presented in Chapters 5 and 6, the confidence intervals were much larger than those shown in Chapters 3 and 4. These large intervals are at the edge of drawing statistically significant conclusions from these experiments. One possible reason was that evaluation in these special VC applications often required domain-specific knowledge. A solution as simple as recruiting professional specialists for evaluation would be, however, expensive or even impossible.

The other solution is to redesign the evaluation protocols. Recently, there has been an increasing backlash on employing MOS tests in speech synthesis evaluation [90,183],

criticizing the lack of detailed information reported in papers [184], its inconsistency between different MOS tests [185], insufficient number of samples used in papers [186], and robustness to anchor systems [187]. There exist some more trustworthy evaluation tests, such as the AB preference test or the MUSHRA test [188], but are not widely adopted due to the increased cost when conducting these systems. Researchers need to reach a consensus on a unified, reliable, and transferrable protocol, to facilitate a better research community.

Acknowledgments

I would like to convey my deepest gratitude to my advisor, Prof. Tomoki Toda of Nagoya University, for everything he has done to support me throughout this journey. He has always been my role model as a researcher, and his expertise and insightful feedback have always pushed me to sharpen my thinking and bring my work to a higher level.

I have been greatly benefitted by the open-source research community since the beginning of my graduate study. It all started with the ESPNet toolkit, which was founded by Prof. Shinji Watanabe from Carnegie Mellon University. He turned out to become one of my most important collaborators, and has always been given me great opportunities and career advice.

I was later fortunate enough to meet the SUPERB team, co-founded by Prof. Hung-Yi Lee, and the S3PRL team, led by Leo Yang, both of whom are from National Taiwan University. It means a lot to work with the people from my alma meter again, and I am always amazed by how creative and knowledgeable they are.

In the last two years of my study, I was lucky enough to work with Prof. Junichi Yamagishi and Dr. Erica Cooper from the National Institute of Informatics, Japan. Together we started a new research direction that turned out to be very impactful to the community. I am often amazed by their rigorousness in research, and I have definitely gained a lot of new experiences along this journey.

Throughout my study, I had the pleasure to work shortly at multiple companies and experience how research is being done in the industries. I would fist like to thank Prof. Hirokazu Kameoka from the NTT Corporation, for the invaluable opportunity to intern at the NTT Communication Science Laboratory.

I could not be more fortunate to have the internship opportunity at the Reality Labs Research (formerly known as Facebook Reality Labs Research). What made it so special was that they allowed me to intern remotely during the pandemic. In addition to the teammates, Alex and Israel, I would like to express my wholehearted thanks to Dejan, who kindly spent his precious night off hours to host meetings with me, in order to tackle the time difference between Taiwan and Pittsburg. He also did everything he could to overcome all the issues in a remote internship. It was truly a pleasure to work with him.

I was then blessed to be able to spend probably the most precious summer in my life to intern at the FAIR (formerly known as Facebook AI Research) New York office. I was lucky enough to be mentored by Peng-Jen Chen – who graduated from the same department at National Taiwan University as I did! I would also like to thank other team members who assisted me throughout my internship: Ben, Justine, Changhan, Hongyu, Yossi, and Ann. I also spent wonderful times chatting with the other two interns, Jeff and Liz. Finally, I would like to thank Juan for reaching out to me in the beginning. Had it not for him, none of this would have happened.

In the last year of my study, I was grateful to work as a student researcher at Google DeepMind. As an AI/ML researcher, it was truly a "dream-come-true" moment to work at Google and witness how research is being conducted by the most gifted frontiers in this research field. My mentor, Dr. Yuma Koizumi, was the kindest person in the world, who supported me in every possible way. I would like to express my deepest

gratitude, and I deeply look forward to working with him again in the future.

I would like to devote my earnest thanks to the staff of Toda Laboratory for their kind assistance. I would also like to convey my thanks to laboratory colleagues for their support, especially Dr. Kazuhiro Kobayashi of TARVO, Inc., and Dr. Tomoki Hayashi of Human Dataware Lab. Co., Ltd., Dr. Yi-Chiao Wu of FAIR, and Prof. Yusuke Yasuda for their aid and advice on my research.

I would like to acknowledge Prof. Hsin-Min Wang of the Institute of Information Science, Academia Sinica Taipei, and Prof. Yu Tsao of the Research Center for Information Technology Innovation, Academia Sinica Taipei, for their patient support and for all of the opportunities I was given to further my research. I would also like to thank my colleagues from the Speech, Language, and Music Processing Lab when I worked as a research assistant in the Institute of Information Science, Academia Sinica Taipei, for the fruitful, inspiring discussion we had and the great memories we shared. I would particularly like to single out Dr. Hsin-Te Hwang, for teaching me the fundamentals of conducting scientific research.

I would especially like to express my humble gratefulness to the Japan Society for the Promotion of Science and NTT Docomo for the financial support they provided throughout my study.

Finally, I would like to express my wholehearted recognition to my family and my girlfriend for their warm company and wise counsel. I would also like to thank the many friends I met in Japan. You enriched my life outside of my research.

References

[1] T. Toda, "Augmented Speech Production based on Real-time Statistical Voice Conversion," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 592–596.

[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid Systems using GMM-based Voice Conversion for Electrolaryngeal Speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[3] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal Speech Enhancement based on One-to-many Eigenvoice Conversion," *IEEE/ACM TASLP*, vol. 22, no. 1, pp. 172–183, 2014.

[4] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems*, vol. E97.D, no. 6, pp. 1429–1437, 2014.

[5] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign Accent Conversion in Computer Assisted Pronunciation Training," *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009.

[6] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign Accent Conversion through Voice Morphing," in *Proc. Interspeech*, 2013, pp. 3077–3081.

[7] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," in *Proc. Interspeech*, 2016, pp. 1632–1636.

[8] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods," in *Proc. Odyssey The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[9] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice Conversion Challenge 2020 - Intra-lingual semi-parallel and cross-lingual voice conversion -," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 80–98.

[10] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[11] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017.

[12] D. Childers, B. Yegnanarayana, and K. Wu, "Voice Conversion: Factors Responsible for Quality," in *Proc. ICASSP*, vol. 10, 1985, pp. 748–751.

[13] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[14] B. Sisman, J. Yamagishi, S. King, and H. Li, "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning," *IEEE/ACM TASLP*, vol. 29, pp. 132–157, 2021.

[15] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice Conversion using Sequence-to-Sequence Learning of Context Posterior Probabilities," in *Proc. Interspeech*, 2017, pp. 1268–1272.

[16] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms," in *Proc. ICASSP*, May 2019, pp. 6805–6809.

[17] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-Sequence Acoustic Modeling for Voice Conversion," *IEEE/ACM TASLP*, vol. 27, no. 3, pp. 631–644, 2019.

[18] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully Convolutional Sequence-to-Sequence Voice Conversion," *IEEE/ACM TASLP*, vol. 28, pp. 1849–1863, 2020.

[19] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 9, 2010, pp. 201–208.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. NIPS*, 2014, pp. 3320–3328.

[22] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition," in *Proc. ICML*, vol. 32, no. 1, 2014, pp. 647–655.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2014, pp. 580–587.

[24] R. Girshick, ""fast r-cnn"," in *Proc. ICCV*, 2015, pp. 1440–1448.

[25] S. Ren, K. He, R. Girshick, and J. Sun, ""faster r-cnn: Towards real-time object detection with region proposal networks"," in *Proc. NIPS*, 2015, pp. 91–99.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ""mask r-cnn"," in *Proc. ICCV*, 2017, pp. 2961–2969.

[28] "Image Style Transfer using Convolutional Neural Networks, author=Gatys, Leon A and Ecker, Alexander S and Bethge, Matthias, booktitle=Proc. CVPR, pages=2414–2423, year=2016."

[29] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc, ECCV*, 2016, pp. 694–711.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. ICML*, vol. 119, 13–18 Jul 2020, pp. 1597–1607.

[31] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," in *Proc. CVPR*, June 2021, pp. 15 750–15 758.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proc. CVPR*, June 2022, pp. 16 000–16 009.

[33] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Jun. 2019, pp. 4171–4186.

[35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019.

[36] "Context-dependent Pre-trained Deep Neural Networks for Large-vocabulary Speech Recognition, author=Dahl, George E and Yu, Dong and Deng, Li and Acero, Alex, journal=IEEE TASLP, volume=20, number=1, pages=30–42, year=2011,."

[37] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[38] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeruIPS*, 2020.

[39] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How Much Can a Bad Teacher Benefit ASR Pre-Training?" in *Proc. ICASSP*, 2021, pp. 6533–6537.

[40] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Proc. NIPS*, vol. 27, 2014, pp. 3104–3112.

[41] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[42] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2015.

[44] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proc. EMNLP*, Sep. 2015, pp. 1412–1421.

[45] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," in *Proc. NIPS*, 2015, pp. 577–585.

[46] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," in *Proc. ICASSP*, 2018, pp. 4784–4788.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[48] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A Comparative Study on Transformer vs RNN in Speech Applications," in *Proc. ASRU*, 2019, pp. 449–456.

[49] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[50] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[51] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition," in *Proc. ICASSP*, 2018, pp. 5884–5888.

[52] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional Sequence to Sequence Learning," in *Proc. ICML*, vol. 70, 2017, pp. 1243–1252.

[53] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural Speech Synthesis with Transformer Network," in *Proc. AAAI*, vol. 33, 2019, pp. 6706–6713.

[54] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE/ACM TASLP*, vol. 14, no. 3, pp. 952–963, 2006.

[55] C.-H. Lee and C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Proc. International Conference on Spoken Language Processing*, 2006.

[56] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *Proc. ICASSP*, 2013, pp. 6905–6909.

[57] R. Aihara, T. Takiguchi, and Y. Ariki, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM TASLP*, vol. 24, no. 7, pp. 1175–1184, 2016.

[58] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, vol. IV, April 2007, p. 2446–24.

[59] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," in *Proc. INTERSPEECH*, 2009, pp. 1623–1626.

[60] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[61] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder," in *Proc. ICASSP*, 2019, pp. 6815–6819.

[62] J. Zhang, Z. Ling, and L. Dai, "Non-Parallel Sequence-to-Sequence Voice Conversion With Disentangled Linguistic and Speaker Representations," *IEEE/ACM TASLP*, vol. 28, pp. 540–552, 2020.

[63] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proc. ICML*, 09–15 Jun 2019, pp. 5210–5219.

[64] D. P. Kingma and M. Welling, "Auto-encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[65] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice Conversion from Non-parallel Corpora using Variational Auto-encoder," in *Proc. APISPA ASC*, 2016, pp. 1–6.

[66] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks," in *Proc. Interspeech*, 2017, pp. 3364–3368.

[67] A. van den Oord, O. Vinyals, and k. Kavukcuoglu, "Neural discrete representation learning," in *NIPS*, 2017, pp. 6306–6315.

[68] A. Tjandra, B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "VQVAE Unsupervised Unit Discovery and Multi-Scale Code2Spec Inverter for Zerospeech Challenge 2019," in *Proc. Interspeech*, 2019, pp. 1118–1122.

[69] D.-Y. Wu and H.-y. Lee, "One-Shot Voice Conversion by Vector Quantization," in *Proc. ICASSP*, 2020, pp. 7734–7738.

[70] J.-C. Chou, C.-C. Yeh, and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Proc. Interspeech*, 2019, pp. 664–668.

[71] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised Speech Decomposition via Triple Information Bottleneck," in *Proc. ICML*, 2020, pp. 7836–7846.

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, vol. 27, 2014.

[73] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, vol. 37, 07–09 Jul 2015, pp. 1530–1538.

[74] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeruIPS*, vol. 33, 2020, pp. 6840–6851.

[75] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[76] W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 160–164.

[77] J.-X. Zhang, L.-J. Liu, Y.-N. Chen, Y.-J. Hu, Y. J., Z.-H. Ling, and L.-R. Dai, "Voice Conversion by Cascading Automatic Speech Recognition and Text-to-Speech Synthesis with Prosody Transfer," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 121–125.

[78] L.-J. Liu, Y.-N. Chen, J.-X. Zhang, Y. Jiang, Y.-J. Hu, Z.-H. Ling, and L.-R. Dai, "Non-Parallel Voice Conversion with Autoregressive Conversion Model and Duration Adjustment," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 126–130.

[79] L. Zheng, J. Tao, Z. Wen, and R. Zhong, "CASIA Voice Conversion System for the Voice Conversion Challenge 2020," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 136–139.

[80] Q. Ma, R. Liu, X. Wen, C. Lu, and X. Chen, "Submission from SRCB for Voice Conversion Challenge 2020," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 131–135.

[81] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Proc. Interspeech*, 2021, pp. 3615–3619.

[82] W.-C. Huang, Y.-C. Wu, T. Hayashi, and T. Toda, "Any-to-One Sequence-to-Sequence Voice Conversion using Self-Supervised Discrete Speech Representations," in *Proc. ICASSP*, 2021, pp. 5944–5948.

[83] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-Y. Lee, and L.-S. Lee, "FragmentVC: Any-to-Any Voice Conversion by End-to-End Extracting and Fusing Fine-Grained Voice Fragments With Attention," in *Proc. ICASSP*, 2021, pp. 5939–5943.

[84] J.-H. Lin, Y. Y. Lin, C.-M. Chien, and H.-Y. Lee, "S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations," in *Proc. Interspeech*, 2021, pp. 836–840.

[85] W.-C. Huang, T. Hayashi, X. Li, S. Watanabe, and T. Toda, "On Prosody Modeling for ASR+ TTS based Voice Conversion," in *Proc. ASRU*, 2021, pp. 642–649.

[86] Y.-H. Peng, C.-H. Hu, A. Kang, H.-S. Lee, P.-Y. Chen, Y. Tsao, and H.-M. Wang, "The Academia Sinica Systems of Voice Conversion for VCC2020," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 180–183.

[87] S. Liu, Y. Cao, S. Kang, N. Hu, X. Liu, D. Su, D. Yu, and H. Meng, "Transferring Source Style in Non-Parallel Voice Conversion," in *Proc. Interspeech*, 2020, pp. 4721–4725.

[88] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, 2016, pp. 1–6.

[89] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet Vocoder with Limited Training Data for Voice Conversion," in *Proc. Interspeech*, 2018, pp. 1983–1987.

[90] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Éva Székely, C. Tånnander, and J. Voße, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 105–110.

[91] ITUT Recommendation, "Methods for Subjective Determination of Transmission Quality," *International Telecommunications Union—Radiocommunication (ITU-T), RITP: Geneva, Switzerland*, 1996.

[92] R. Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.

[93] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. 99, pp. 1877–1884, 2016.

[94] R. K. Das, T. Kinnunen, W.-C. Huang, Z.-H. Ling, J. Yamagishi, Z. Yi, X. Tian, and T. Toda, "Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions," in *Proc. Joint Workshop for the BC and VCC 2020*, 2020, pp. 99–120.

[95] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014, pp. 4052–4056.

[96] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[97] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, p. 369–376.

[98] J. Kominek and A. W. Black, "The CMU ARCTIC Speech Databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[99] Munich Artificial Intelligence Laboratories GmbH, "The M-AILABS speech dataset," 2019, accessed 30 November 2019. [Online]. Available: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

[100] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[101] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit," in *Proc. ICASSP*, 2020, pp. 7654–7658.

[102] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[103] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large batch optimization for deep learning: Training bert in 76 minutes," in *Proc. ICLR*, 2020.

[104] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[105] B. Naderi and R. Cutler, "An Open Source Implementation of ITU-T Recommendation P.808 with Validation," in *Proc. Interspeech*, 2020, pp. 2862–2866.

[106] ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*, Std., 2018.

[107] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[108] Y.-A. Chung and J. Glass, "Generative Pre-Training for Speech with Autoregressive Predictive Coding," in *Proc. ICASSP*, 2020, pp. 3497–3501.

[109] Y.-A. Chung, H. Tang, and J. Glass, "Vector-Quantized Autoregressive Predictive Coding," in *Proc. Interspeech*, 2020, pp. 3760–3764.

[110] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders," in *Proc. ICASSP*, 2020.

[111] A. T. Liu, S.-W. Li, and H.-y. Lee, "TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech," *IEEE/ACM TASLP*, vol. 29, pp. 2351–2366, 2021.

[112] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies," in *Proc. Interspeech*, 2021, pp. 3730–3734.

[113] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[114] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised Pretraining Transfers Well Across Languages," in *Proc. ICASSP*, 2020, pp. 7414–7418.

[115] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.

[116] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations," in *Proc. ICLR*, 2020.

[117] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *Proc. ICASSP*, 2020, pp. 6989–6993.

[118] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. ICASSP*, 2020, pp. 7669–7673.

[119] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Proc. Interspeech*, 2021, pp. 1509–1513.

[120] W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, S. Watanabe, and T. Toda, "S3PRL-VC: Open-Source Voice Conversion Framework with Self-Supervised Speech Representations," in *Proc. ICASSP*, 2022, pp. 6552–6556.

[121] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. t. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Y. Lee, "SUPERB: Speech processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[122] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, vol. 162, 2022, pp. 2709–2720.

[123] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[124] X. Wang, S. Takaki, and J. Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4895–4899.

[125] X. Wang, S. Takaki, and J. Yamagishi, "Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis," *IEEE/ACM TASLP*, vol. 26, no. 8, pp. 1406–1419, 2018.

[126] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[127] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[128] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised Speech Recognition," in *Proc. NeulIPS*, 2021.

[129] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[130] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[131] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[132] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, 2017, pp. 712–718.

[133] S. Liu, Y. Cao, X. Wu, L. Sun, X. Liu, and H. Meng, "Jointly Trained Conversion Model and WaveNet Vocoder for Non-Parallel Voice Conversion Using Mel-Spectrograms and Phonetic Posteriorgrams," in *Proc. Interspeech*, 2019, pp. 714–718.

[134] W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Refined WaveNet Vocoder for Variational Autoencoder Based Voice Conversion," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[135] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities," in *Proc. ACL (Volume 1: Long Papers)*, May 2022, pp. 8479–8492.

[136] R. D. Kent, "Research on speech motor control and its disorders: A review and prospective," *Journal of Communication disorders*, vol. 33, no. 5, pp. 391–428, 2000.

[137] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech and Language*, vol. 27, no. 6, pp. 1163 – 1177, 2013.

[138] A. B. Kain, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743 – 759, 2007.

[139] S. Fu, P. Li, Y. Lai, C. Yang, L. Hsieh, and Y. Tsao, "Joint dictionary learning-based non-negative matrix factorization for voice conversion to improve speech intelligibility after oral surgery," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2584–2594, 2017.

[140] Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, "Voice conversion for persons with amyotrophic lateral sclerosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2942–2949, 2020.

[141] C.-Y. Chen, W.-Z. Zheng, S.-S. Wang, Y. Tsao, P.-C. Li, and Y.-H. Lai, "Enhancing Intelligibility of Dysarthric Speech Using Gated Convolutional-Based Voice Conversion System," in *Proc. Interspeech*, 2020, pp. 4686–4690.

[142] D. Wang, J. Yu, X. Wu, S. Liu, L. Sun, X. Liu, and H. Meng, "End-To-End Voice Conversion Via Cross-Modal Knowledge Distillation for Dysarthric Speech Reconstruction," in *Proc. ICASSP*, 2020, pp. 7744–7748.

[143] M. Purohit, M. Patel, H. Malaviya, A. Patil, M. Parmar, N. Shah, S. Doshi, and H. A. Patil, "Intelligibility improvement of dysarthric speech using mmse discogan," in *Proc. SPCOM*, 2020, pp. 1–5.

[144] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, "Adversarial Data Augmentation for Disordered Speech Recognition," in *Proc. Interspeech*, 2021, pp. 4803–4807.

[145] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary," in *Proc. ICASSP*, 2021, pp. 6428–6432.

[146] B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. Magimai-Doss, "An Objective Evaluation Framework for Pathological Speech Synthesis," in *Speech Communication; 14th ITG Conference*, 2021, pp. 1–5.

[147] M. Illa, B. M. Halpern, R. van Son, L. Moro-Velazquez, and O. Scharenborg, "Pathological Voice Adaptation with Autoencoder-based Voice Conversion," in *Proc. ISCA Speech Synthesis Workshop*, 2021, pp. 19–24.

[148] T. V. H. and M. A., "Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 140–144.

[149] K. Kobayashi, W.-C. Huang, Y.-C. Wu, P. L. Tobing, T. Hayashi, and T. Toda, "crank: An Open-Source Software for Nonparallel Voice Conversion Based on Vector-Quantized Variational Autoencoder," in *Proc. ICASSP*, 2021.

[150] H. Kameoka, W. C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice transformer network," *IEEE/ACM TASLP*, vol. 29, pp. 656–670, 2021.

[151] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[152] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech*, 2019, pp. 4115–4119.

[153] M.-W. Huang, "Development of taiwan mandarin hearing in noise test," Master's thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health Science, 2005.

[154] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved Lite Audio-Visual Speech Enhancement," *IEEE/ACM TASLP*, vol. 30, pp. 1345–1359, 2022.

[155] C.-Y. Tseng, Y.-C. Cheng, and C. Chang, "Sinica COSPRO and Toolkit - Corpora and platform of Mandarin Chinese fluent speech," in *Proc. O-COCOSDA*, 2005, pp. 23–28.

[156] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, 2017, pp. 1–5.

[157] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[158] M. Purohit, M. Patel, H. Malaviya, A. Patil, M. Parmar, N. Shah, S. Doshi, and H. A. Patil, "Intelligibility Improvement of Dysarthric Speech using MMSE

DiscoGAN," in *Proc. International Conference on Signal Processing and Communications (SPCOM)*, 2020, pp. 1–5.

[159] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, "Improving Sequence-to-sequence Voice Conversion by Adding Text-supervision," *Proc. ICASSP*, pp. 6785–6789, 2019.

[160] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-Autoregressive Sequence-To-Sequence Voice Conversion," in *Proc. ICASSP*, 2021, pp. 7068–7072.

[161] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors – in search of the golden speaker," *Speech Communication*, vol. 37, no. 3, pp. 161–173, 2002.

[162] "Golden Speaker Builder – An Interactive Tool for Pronunciation Training, journal = Speech Communication, volume = 115, pages = 51-66, year = 2019, author = Shaojin Ding and Christopher Liberatore and Sinem Sonsaat and Ivana Lučić and Alif Silpachai and Guanlong Zhao and Evgeny Chukharev-Hudilainen and John Levis and Ricardo Gutierrez-Osuna,."

[163] O. Turk and L. M. Arslan, "Subband Based Voice Conversion," in *Proc. 7th International Conference on Spoken Language Processing*, 2002, pp. 289–292.

[164] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 433–446, 2015.

[165] D. Felps, C. Geng, and R. Gutierrez-Osuna, "Foreign Accent Conversion Through Concatenative Synthesis in the Articulatory Domain," *IEEE TASLP*, vol. 20, no. 8, pp. 2301–2312, 2012.

[166] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech and Language*, vol. 36, pp. 260–273, 2016.

[167] S. Aryal and R. Gutierrez-Osuna, "Can voice conversion be used to reduce non-native accents?" in *Proc. ICASSP*, 2014, pp. 7879–7883.

[168] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," in *Proc. ICASSP*, 2018, pp. 5314–5318.

[169] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams," in *Proc. Interspeech*, 2019, pp. 2843–2847.

[170] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "End-to-end accent conversion without using native utterances," in *Proc. ICASSP*, 2020, pp. 6289–6293.

[171] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM TASLP*, vol. 29, pp. 2367–2381, 2021.

[172] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-Shot Foreign Accent Conversion without a Native Reference," in *Proc. Interspeech*, 2022, pp. 4920–4924.

[173] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A Non-native English Speech Corpus," in *Proc. Interspeech*, 2018, pp. 2783—-2787.

[174] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[175] M. J. Munro and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language learning*, vol. 45, no. 1, pp. 73–97, 1995.

[176] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright, "When is it better to compare than to score?" *arXiv preprint arXiv:1406.6618*, 2014.

[177] C. Cucchiarini, H. Van hamme, O. van Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," in *Proc. LREC*, May 2006.

[178] P. L. Tobing and T. Toda, "Low-latency real-time non-parallel voice conversion based on cyclic variational autoencoder and multiband WaveRNN with data-driven linear prediction," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 142–147.

[179] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, "Real-Time, Full-Band, Online DNN-Based Voice Conversion System Using a Single CPU," in *Proc. Interspeech*, 2020, pp. 1021–1022.

[180] Y.-Y. Ding, L.-J. Liu, Y. Hu, and Z.-H. Ling, "A Study on Low-Latency Recognition-Synthesis-Based Any-to-One Voice Conversion," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 455–460.

[181] T. Hayashi, K. Kobayashi, and T. Toda, "An Investigation of Streaming Non-Autoregressive sequence-to-sequence Voice Conversion," in *Proc. ICASSP*, 2022, pp. 6802–6806.

[182] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, Y. Yasuda, and T. Toda, "The Singing Voice Conversion Challenge 2023," in *to appear in Proc. ASRU*, 2023.

[183] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech and Language*, vol. 84, p. 101577, 2024.

[184] C.-H. Chiang, W.-P. Huang, and H. yi Lee, "Why We Should Report the Details in Subjective Evaluation of TTS More Rigorously," in *Proc. Interspeech*, 2023, pp. 5551–5555.

[185] S. Le Maguer, S. King, and N. Harte, "Back to the Future: Extending the Blizzard Challenge 2013," in *Proc. Interspeech*, 2022, pp. 2378–2382.

[186] Y. Yasuda and T. Toda, "Analysis of Mean Opinion Scores in Subjective Evaluation of Synthetic Speech Based on Tail Probabilities," in *Proc. Interspeech*, 2023, pp. 5491–5495.

[187] E. Cooper and J. Yamagishi, "Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech," in *Proc. Interspeech*, 2023, pp. 1104–1108.

[188] "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

List of Publications

Journal Papers

1. W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, T. Toda, "Pretraining Techniques for Sequence-to-sequence Voice Conversion," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 745–755, 2021.

2. W.-C. Huang, S.-W. Yang, T. Hayashi, T. Toda, "A Comparative Study of Self-Supervised Speech Representation Based Voice Conversion," IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1308–1318, 2022.

3. W.-C. Huang, H. Luo, H.-T. Hwang, C.-C. Lo, Y.-H. Peng, Y. Tsao, H.-M. Wang, "Unsupervised Representation Disentanglement Using Cross Domain Features and Adversarial Learning in Variational Autoencoder Based Voice Conversion," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 4, no. 4, pp. 468–479, 2020.

4. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf,

J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: a large-scale public database of synthesized, converted and replayed speech," Computer Speech and Language, vol. 64, p. 101114, 2020

5. H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, T. Toda, "Many-to-many voice transformer network," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 29, pp. 656-670, Jan. 2021.

International Conferences and Workshops

1. W.-C. Huang, Y.-C. Wu, H.-T. Hwang, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, H.-M. Wang. "Refined WaveNet vocoder for variational autoencoder based voice conversion," Proc. EUSIPCO, 2019, pp. 1–5.

2. W.-C. Huang, Y.-C. Wu, C.-C. Lo, P. L. Tobing, T. Hayashi, K. Kobayashi, T. Toda, Y. Tsao, and H.-M. Wang, "Investigation of F0 Conditioning and Fully Convolutional Networks in Variational Autoencoder Based Voice Conversion,", Proc. Interspeech, 2019, pp. 709-713

3. W.-C. Huang, Y.-C. Wu, K. Kobayashi, Y.-H. Peng, H.-T. Hwang, P. L. Tobing, T. Toda, Y. Tsao, and H.-M. Wang, "Generalization of Spectrum Differential based Direct Waveform Modification for Voice Conversion," Proc. ISCA Speech Synthesis Workshop (SSW), 2019, pp. 57-62

4. W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining," Proc. Interspeech, 2020, pp. 4676–4680

5. W.-C. Huang, P. L. Tobing, Y.-C. Wu, K. Kobayashi, and T. Toda, "The NU Voice Conversion System for the Voice Conversion Challenge 2020: On the Effectiveness of Sequence-to-sequence Models and Autoregressive Neural Vocoders," Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020, pp. 165-169

6. W.-C. Huang, T. Hayashi, S. Watanabe, and T. Toda, "The Sequence-to-Sequence Baseline for the Voice Conversion Challenge 2020: Cascading ASR and TTS," Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020, pp. 160-164

7. W.-C. Huang, T. Hayashi, Y.-C. Wu, and T. Toda, "Any-to-One Sequence-to-Sequence Voice Conversion using Self-Supervised Discrete Speech Representations," in Proc. ICASSP, 2021, pp. 5944–5948

8. W.-C. Huang, C.-H. Wu, S.-B. Luo, K.-Y. Chen, H.-M. Wang, and T. Toda, "Speech recognition by simply fine-tuning bert," in Proc. ICASSP, 2021, pp. 7343–7347

9. W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, and T. Toda, "A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion," in Proc. Interspeech, 2021, pp. 1329–1333

10. W.-C. Huang, T. Hayashi, X. Li, S. Watanabe, and T. Toda, "On Prosody Modeling for ASR+TTS based Voice Conversion," in Proc. ASRU, 2021, pp. 642-649

11. W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards Identity Preserving Normal to Dysarthric Voice Conversion," in Proc. ICASSP, 2022, pp. 6672-6676

12. W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, “LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech,” in Proc. ICASSP, 2022, pp. 896-900

13. W.-C. Huang, S.-W. Yang, T. Hayashi, H.-Y. Lee, S. Watanabe, and T. Toda, “S3PRL-VC: Open-source Voice Conversion Framework with Self-supervised Speech Representations,” in Proc. ICASSP, 2022, pp. 6552-6556

14. W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, J. Yamagishi, and T. Toda, “The VoiceMOS Challenge 2022,” in Proc. Interspeech, 2022, pp. 4536-4540

15. W.-C. Huang, D. Markovic, A. Richard, I. D. Gebru, and A. Menon, “End-to-End Binaural Speech Synthe- sis,” in Proc. Interspeech, 2022, pp. 1218-1222

16. W.-C. Huang, B. Peloquin, J. Kao, C. Wang, H. Gong, E. Salesky, Y. Adi, A. Lee, and P.-J. Chen, “A Holistic Cascade System, Benchmark, and Human Evaluation Protocol for Expressive Speech-to-Speech Translation,” in Proc. ICASSP, 2023, pp. 1–5

17. W.-C. Huang, T. Toda, “Evaluating Methods for Ground-Truth-Free Foreign Accent Conversion,” in Proc. APSIPA ASC, 2023, pp. 1136-1141

18. W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, Y. Yasuda, and T. Toda “The Singing Voice Conversion Challenge 2023,” to appear in Proc. ASRU, 2023, 8 pages

19. W.-C. Huang, E. Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi, “The VoiceMOS Challenge 2023: Zero-shot Subjective Speech Quality Prediction for Multiple Domains,” to appear in Proc. ASRU, 2023, 8 pages

20. C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep Learning based Objective Assessment for Voice Conversion," in Proc. Interspeech, 2019, pp. 1541-1545

21. Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda, "Voice Conversion Challenge 2020 – Intra-lingual semi-parallel and cross-lingual voice conversion –," in Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020, pp. 80-98

22. R. K. Das, T. Kinnunen, W.-C. Huang, Z.-H. Ling, J. Yamagishi, Z. Yi, X. Tian, and T. Toda, "Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions," in Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020, pp. 99-120

23. Y.-W. Chen, K.-H. Hung, S.-Y. Chuang, J. Sherman, W.-C. Huang, X. Lu, and Y. Tsao, "Ema2s: An end-to-end multimodal articulatory-to-speech system," in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), 2021, pp. 1–5

24. T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in Proc. ICASSP, 2021, pp. 7068–7072

25. K. Kobayashi, W.-C. Huang, Y.-C. Wu, P. L. Tobing, T. Hayashi, and T. Toda, "CRANK: an Open-Source Software for Nonparallel Voice Conversion based on Vetor-Quantized Variational Autoencoder," in Proc. ICASSP, 2021, pp. 5934–5938

26. Y.-C. Wu, C.-H. Hu, H.-S. Lee, Y.-H. Peng, W.-C. Huang, Y. Tsao, H.-M. Wang, and T. Toda, "Rela- tional Data Selection for Data Augmentation of

Speaker-dependent Multi-band MelGAN Vocoder," in Proc. Interspeech, 2021, pp. 3630–3634

27. D. Ma, W.-C. Huang, and T. Toda, "Investigation of Text-to-speech-based Synthetic Parallel Data for Sequence-to-sequence Non-parallel Voice Conversion," in Proc. APSIPA ASC, 2021, pp. 870-877

28. C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy Voice Conversion Framework with Denoising Model," in Proc. APSIPA ASC, 2021, pp. 814-820

29. Y.-S. Liou, W.-C. Huang, M.-C. Yen, S.-W. Tsai, Y.-H. Peng, T. Toda, Y. Tsao, and H.-M. Wang, "Time Alignment using Lip Images for Frame-based Electrolaryngeal Voice Conversion," in Proc. APSIPA ASC, 2021, pp. 1234-1238

30. M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. Jang, and H.-M. Wang, "Mandarin Electrolaryngeal Speech Voice Conversion with Sequence-to-Sequence Modeling," in Proc. ASRU, 2021, pp. 650-657

31. C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Direct Noisy Speech Modeling for Noisy-to-noisy Voice Conversion," in Proc. ICASSP, 2022, pp.6787-6791

32. E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in Proc. ICASSP, 2022, pp. 8442-8446

33. H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi et al., "SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities,"

in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8479–8492

34. L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition," in Proc. Interspeech, 2022, pp. 41-45

35. L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Intermediate Fine-Tuning Using Imperfect Synthetic Speech for Improving Electrolaryngeal Speech Recognition," in Proc. ICASSP, 2023, 5 pages

Awards

1. Best Student Paper Award, The 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018

2. Best Paper Award, The 13th Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2021

3. 16th Student Journal Paper Award, IEEE SPS Japan, Dec. 2022

4. Outstanding Graduate Student Award, from Nagoya University, Japan, June 2023

Organizing Committee

1. Voice Conversion Challenge 2020, Mar. 2023 to Oct. 2023

2. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, Oct. 2020

3. VoiceMOS Challenge 2022, Nov. 2021 to Mar. 2022

4. Singing Voice Conversion Challenge 2023, Jan. 2023 to June 2023

5. VoiceMOS Challenge 2023, Mar. 2023 to June 2023