Training Techniques of Sequence-to-sequence Voice Conversion for Electrolaryngeal Speech Enhancement

Ding MA

Abstract

Communication barriers faced by individuals with disabilities remain a significant societal concern. One representative group is laryngectomees, who depends on an external electromechanical device to substitute for the vocal folds, producing ElectroLaryngeal (EL) speech for verbal communication. While such devices restore the basic ability to speak, EL speech is markedly degraded compared with normal speech in terms of quality, naturalness, and intelligibility, severely limiting the quality of life of EL speakers. To address this issue, a growing body of research has explored Voice Conversion (VC) for ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP). Early EL2SP systems were based on frame-wise VC, which rigidly converted parallel acoustic frames. As such, these methods lacked flexibility in temporal modeling, causing the converted speech to inherit the duration patterns of EL inputs and thus sound unnatural. With the advancement of deep neural networks, sequence-to-sequence (seq2seq) VC has emerged as a promising alternative for EL2SP. It enables temporal structure conversion by modeling variable-length EL and normal speech sequences, leading to more natural prosody and smoother rhythm of converted speech.

This thesis aims to address three critical challenges in applying seq2seq modeling for EL2SP: (1) Contradiction between limited EL2SP dataset and huge data requirement of seq2seq modeling, (2) Lack of robustness to real-world interferences, and (3) Cumulative errors in internal mapping function. For the first challenge, the focus is placed on two aspects: (1) Obtaining more easily accessible training data, and (2) Developing methods that enhance cross-domain transferability and stabilize performance. For the

second challenge, attention is directed toward constructing a unified EL2SP system capable of flexibly handling various interfered conditions. To address the final challenge, efforts are concentrated on acquiring more accurate representations. Based on these considerations, a series of efficient training techniques are proposed within the main concepts of data augmentation and transfer learning.

First, to address the data scarcity issue in EL2SP, novel multi-stage pretraining and fine-tuning techniques are proposed using easily accessible, low-quality Synthetic Data (SD). Aside from transferring Text-To-Speech (TTS) pretraining to seq2seq VC based on a normal TTS corpus, an encoder adaptation training is designed to minimize the domain-shift gap between upstream normal and downstream EL2SP data. Then, a two-stage fine-tuning scheme is introduced to continuously enhance transfer learning and stabilize the final performance. Moreover, by leveraging low-quality SD as data augmentation materials for both encoder adaptation and fine-tuning processes, the proposed methods largely reduce data requirements and broaden the scope of practical application. Experimental results demonstrate the effectiveness of the proposed stages in the proposed method, and the resulting systems dramatically outperform a conventional method based on direct pretraining–fine-tuning pipeline.

Second, to address the model's inadaptability for real-world scenarios where EL speech is interfered with background noise and reverberation, a training method with a multi-source learning paradigm is proposed. This method requires only a single framework rather than using any extra speech enhancement module to tackle interferences. In particular, a two-stage fine-tuning is employed in a many-to-one style, leveraging pseudo-noisy and reverberant EL speech data generated from limited clean data. Different system designs are evaluated and the baselines are established including models trained on clean data or combined with external SE modules. Comparative experiments demonstrate that the proposed method consistently outperforms these baselines, effectively handling both clean and noisy conditions.

Finally, to address mapping inaccuracy in seq2seq modeling for EL2SP, a novel

representation learning framework is proposed for integrating paired EL speech and text representations. A joint network training first incorporates the text information to enrich intermediate representations. Next, an autoencoder-style reconstruction training is proposed to align the speech encoder with these enriched representations and enhance the mapping between speech encoder and decoder, finalizing the better-performing EL2SP model without added complexity. Around this method, the use of data augmentation in joint network training, and a preliminary loss design in reconstruction training, are further explored, respectively. Experimental results demonstrate that the proposed method with data augmentation, significantly outperforms the baselines that learn speech-only representations. Meanwhile, the deeper loss design contributes to a basic optimization for the final performance.

To summarize, rather than altering the seq2seq architecture or relying on expensive resources, this thesis focuses on designing effective training strategies to cost-efficiently enhance data efficiency, robustness, and mapping accuracy, thereby extending the feasibility of seq2seq EL2SP for practical use. The experimental findings and discussions also open up new research directions for future work.

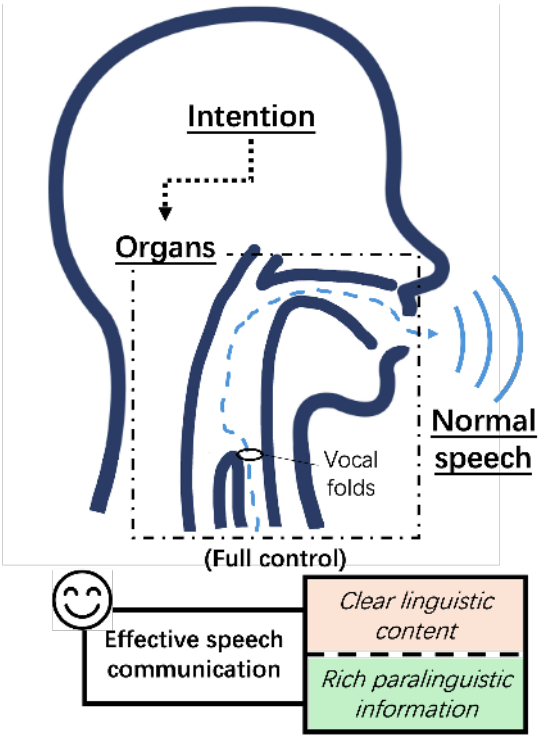
Contents

1 Introduction

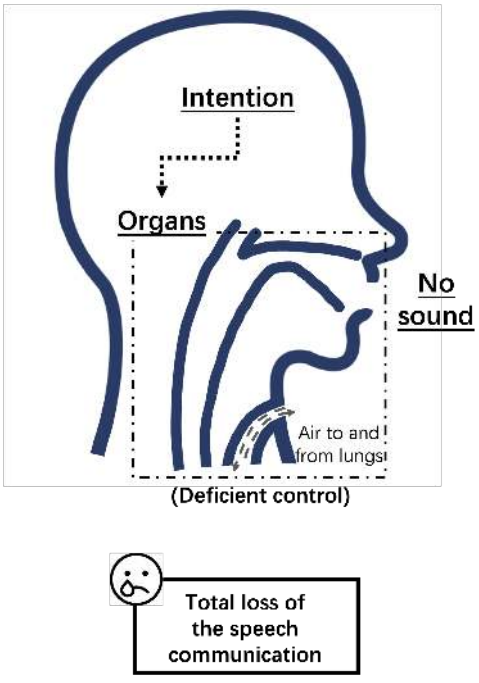
1.1 General Background

Speech communication plays an important role in human-to-human interactions, during which speech does not manifest in isolation. Rather, it is often shaped by the context and intentions of the speaker involved, serving particular purpose at corresponding moments and occasions. Specifically, when people talk, speech does not merely carry words; it also conveys a wide range of information that makes communication natural and effective. In general, the primary functions of speech signals from a healthy speaker comprise: (1) Clear linguistic content, which transmits the actual words and grammar, (2) Paralinguistic information [1], such as accent, prosody, and emotion, and (3) Speaker-dependent nonlinguistic information, such as speaker identity. While linguistic content conveys the intended message, paralinguistic and nonlinguistic information are also essential for achieving proper and natural expression in human speech communication. Collectively, the three aspects allow humans to exchange both accurate meaning and personal characteristics.

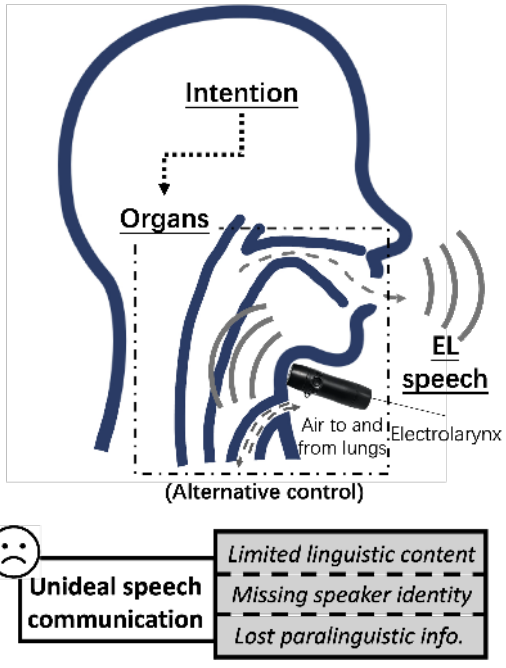
The production of such rich speech is determined by the physical mechanisms of the human body. The vocal folds are the most important organ in speaking, which vibrates the airflow from trachea to produce source excitation sounds for speech generation. These excitation sounds are then modulated by the specific articulatory configurations of the vocal tract, including the tongue, lips, and oral cavity, along with their resonance characteristics, to generate speech we hear in everyday conversations



(a) Normal speaker.



(b) After laryngectomy, without aid.



(c) Leveraging electrolarynx.

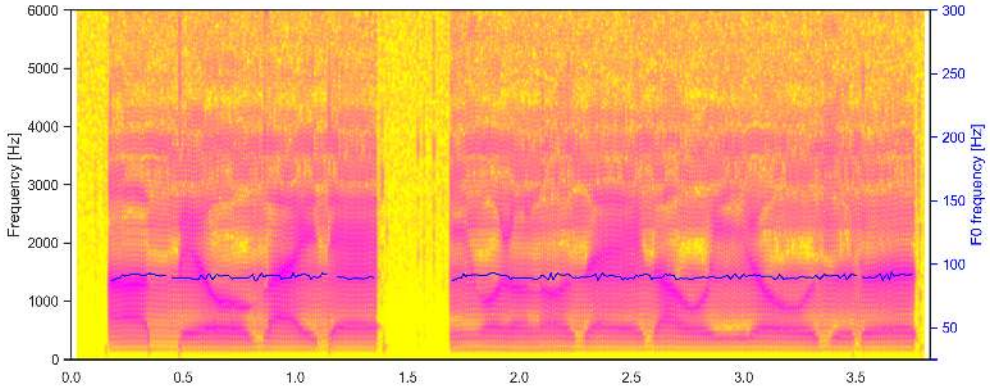
Figure 1.1: Illustration of speech information retention.

[2], as shown in Figure 1.1(a). Based on this, individuals who suffer from the damage or lost control for their vocal folds cannot completely convey both linguistic and paralinguistic information, and thus face severe communication barriers and life distresses [3], [4].

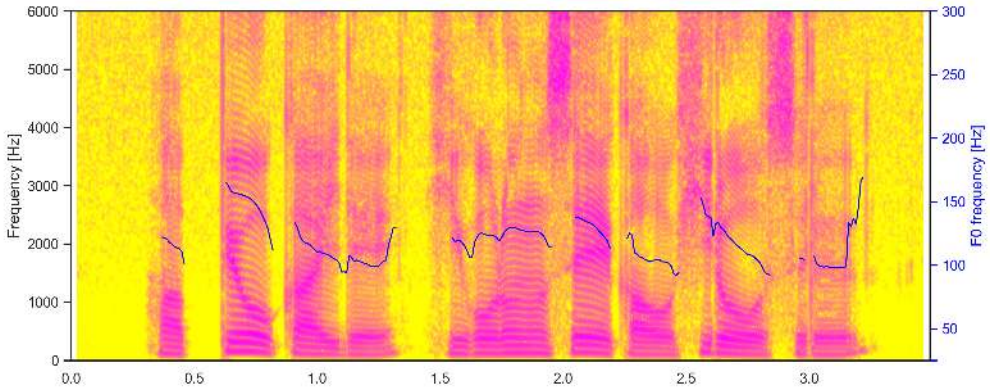
A representative case is patients who undergo total laryngectomy, a surgical removal of the larynx, often as a treatment for laryngeal cancer [5]. Although patients retain the knowledge of how to speak, this surgery results in the complete loss of their original sound source organs, including vocal folds, preventing them produce speech anymore, as demonstrated in Figure 1.1(b). Over the years, doctors and researchers have developed several alternative methods for laryngectomees to help compensate for this loss and rehabilitate their speaking ability.

One method is esophageal speech. Patients push or swallow air into the esophagus, and then create sound by using esophageal walls as a vibration source and pushing air back up to the mouth while shaping the words. However, esophageal speech is not a general option due to following fundamental constraints: (1) Its feasibility is determined by the extent of laryngectomy, and not all patients can successfully master it, (2) It is very hard to learn, necessitating many months of intensive training with a speech therapist, and (3) The acquired air to the esophagus is insufficient, resulting in discontinuous and short utterances. Another method is tracheoesophageal speech. It involves a surgical procedure called TracheoEsophageal Puncture (TEP) so as to implant a one-way valve that redirects pulmonary air into the esophagus. TEP yields more fluent speech than esophageal speech, but this approach still requires long implementation period, involving additional surgery, continuous medical follow-ups and complex care, imposing further physical and psychological burdens on patients.

In contrast to the aforementioned methods, ElectroLaryngeal (EL) speech [6]–[8] requires much less time cost and fewer risks for patients. As shown in Figure 1.1(c), EL speech is an artificial speech produced through a portable device named electrolarynx, which is placed against the throat to simulate the vibrations of the vocal folds exter-



(a) ElectroLaryngeal (EL) speech



(b) Normal speech

Figure 1.2: Visualization of the mel spectrograms and F0 patterns of (a) ElectroLaryngeal (EL) speech and (b) normal speech uttering the same sentence.

nally. The source excitation sounds it generates are conducted into the oral cavity and articulated to form the EL speech. Remarkably, many laryngectomees can begin using electrolarynx to produce EL speech only 3–5 days after surgery. For these reasons, EL speech has become the most convenient and widely used method among patients. However, due to the inherent limitations in its artificial sound generation mechanism, EL speech sounds monotonous, robotic, and unnatural, lacking linguistic clarity, paralinguistic expressiveness, and speaker identity compared with normal speech.

To further highlight these shortcomings, Figure 1.2 shows the substantial gap between the features of EL speech and normal speech, mainly due to two significant is-

sues. First, the mechanically generated excitation signals cannot replicate variable fundamental frequency (F0) contours of a healthy voice [9], including changes in pitches, intonation, and some voiced/unvoiced sounds [10]. This causes the loss of essential information such as prosody and emotional expression. Meanwhile, the absence of intonation changes and distinctions in voiced/unvoiced sounds make EL speech challenging to comprehend [11]. Furthermore, the temporal structure (i.e., duration) of EL speech also tends to significantly differ from that of normal speech, as reflected in Figure 1.2. Moreover, the high-energy excitation signals are inevitably emitted outside, accompanied by intense noise, which cause poor quality and degraded intelligibility of EL speech. These issues make EL speech less clear than normal speech and may cause discomfort to both patients and listeners.

In summary, given the convenience and wide applicability of EL speech compared with other speaking methods, addressing its limitations to improve the impaired quality of life for patients forms the central motivation of this thesis.

1.2 Voice Conversion (VC)

What if one can enhance the impaired functions of laryngectomees, making them comparable to those of normal speakers? This is where *Voice Conversion* (VC) can play a role. It is a methodology that involves converting the voice of one speaker into that of another while preserving the linguistic contents [12]–[15]. Therefore, leveraging VC to realize EL-speech-to-normal-SPeech conversion (EL2SP), can be viewed as a promising approach to overcome the physical constraints associated with speech communication in larygectomees.

Data deficiency has long posed a significant challenge in VC, as almost all VC methods rely on data-driven approaches. From early statistical methods to recent models based on Deep Neural Networks (DNNs), the demand for data has continuously increased. However, recording datasets for VC is both time-consuming and labor-

intensive, especially since VC tasks require collecting both source and target speech data. In contrast, data collection for other speech processing tasks, such as Text-To-Speech (TTS) or Automatic Speech Recognition (ASR), is relatively easier, because TTS and ASR are formulated as learning from paired *speech* and *text* data, rather than paired *speech* and *speech* data as in VC. Although mapping paired *speech* and *text* data is more complex and therefore requires a larger amount of training data (e.g., LJSpeech [16] with approximately 10 hours, VCTK [17] with around 44 hours, and LaboroTV [18] with about 2,000 hours), obtaining text corresponding to speech is much easier than obtaining another corresponding speech, as the latter incurs higher recording cost. As a result, VC datasets are far more limited. For instance, the available VC datasets, involving normal-to-normal conversion tasks or challenges, are often allocated only around 1 hour [19], or even just a few minutes [20]–[22], for each source and target speaker, further underscoring this issue.

For the EL2SP task addressed in this thesis, data deficiency is even more severe. As physiological functions of patients are more fragile, it is extremely impractical for them to undergo lengthy recording periods. In addition, obtaining healthy speech samples from patients is also challenging in many cases. Therefore, data deficiency in EL2SP involves not only the scarcity of training data but also the lack of fully *parallel*[1] source–target data pairs, often referred to as *semi-parallel corpus*. This situation exacerbates the domain gap between EL and normal speech, which can be regarded as a sub-issue of data deficiency. Specifically, the significant feature differences between EL and normal speech complicate speech mapping and conversion to VC models, and the limited data further hinders the model in establishing an effective mapping function.

Thus, the primary focus of this thesis is to solve the data deficiency in EL2SP. To clearly outline this challenge, we will review the mainstream methods and frameworks to EL2SP in the following sections, and discuss their potential limitations.

[1] *parallel* represents a set of utterance pairs from the source and target with identical linguistical contents.

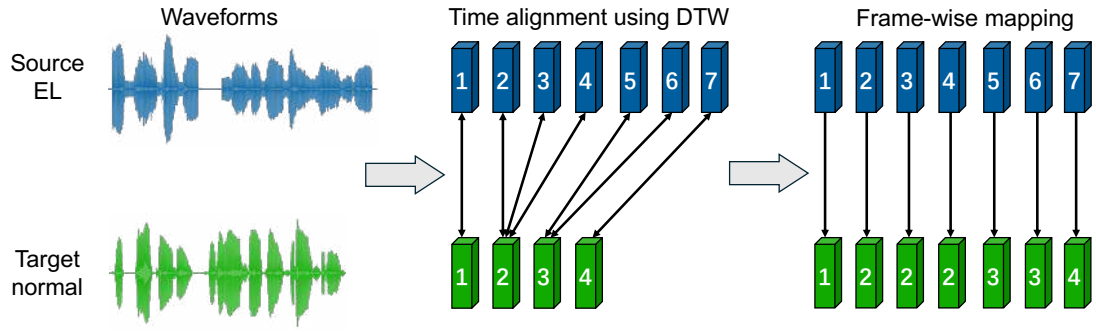


Figure 1.3: Illustration of the frame-wise voice conversion framework for EL-speech-to-normal-SPeech conversion (EL2SP), sharing the same-length source EL and target normal speech sequences for mapping process.

1.2.1 Frame-wise VC

A widely-used approach to EL2SP relies on conventional statistical VC models [2], [23]–[28] that adopt an *analysis–mapping–reconstruction* pipeline using a *parallel corpus*. In this approach, a mapping module converts the features of the source EL speech, which are first decomposed by a speech analyzer (e.g., WORLD [29] or STRAIGHT [30]), to match the target normal speech. A synthesizer subsequently reconstructs the converted time-domain speech signals. Throughout this process, the performance of EL2SP is predominantly determined by the mapping module, which typically employs a *frame-to-frame* mapping paradigm, such as using a feature-based Dynamic Time Warping (DTW) algorithm [31]. As illustrated in Figure 1.3, following the assumption that frames with the same linguistic information in a parallel corpus are close in distance within the feature space (although this may not hold true in practice) [32], DTW explicitly aligns the acoustic features of each frame of the EL and normal speech pair, ensuring that the source and target acoustic feature sequences are of the same length for the final time alignment.

However, this frame-wise paradigm leads to a fundamental problem: It imposes the temporal structure of the converted normal speech to be identical with that of the EL speech, which contradicts the properties of EL and normal speech discussed in Section 1.1. This problem prevents the paradigm to capture the significant differences

in time-variant characteristics, such as rhythm and duration, between the two. As a result, the performance of the conventional EL2SP approach remains limited, particularly in capturing long-term dependencies and converting prosodic features, which degrades speech quality, naturalness, and similarity.

1.2.2 Sequence-to-sequence (seq2seq) VC

Over the past decade, the growth of DNNs has inspired researchers to propose sequence-to-sequence (seq2seq) models [33] to address the aforementioned shortcomings for aligment of conventional ones. Seq2seq models, originally developed for machine translation tasks [34], [35], are designed to map an entire input sequence to an output sequence of potentially different length in a DNN-based end-to-end manner, thereby eliminating the rigid temporal constraints of *frame-to-frame* mapping. This capability makes seq2seq models particularly well-suited for VC applications that require flexible temporal alignment and complex transformations. As shown in Figure 1.4, seq2seq VC models employ an encoder–decoder framework with an attention mechanism to implicitly perform the mapping process [36]. The encoder extracts high-dimensional *intermediate representations*, which primarily contain the pure linguistic information, from the source speech features. The decoder then uses the attention mechanism to focus on different parts of these encoded representations, rather than relying solely on fixed-length context vectors, and combines them with the target characteristics to reconstruct the converted speech features. Therefore, seq2seq VC models are able to automatically determine the output duration of converted speech, as well as to capture long-term dependencies such as prosody, suprasegmental characteristics of F0, and speaker identity [37].

As seq2seq VC models are structurally more advanced, they naturally hold promising prospects to more effectively address the challenges in EL2SP. Several studies [38], [39] have shown that seq2seq VC can outperform frame-based models in normal-to-normal tasks regarding both naturalness and conversion similarity. Coincidentally, the

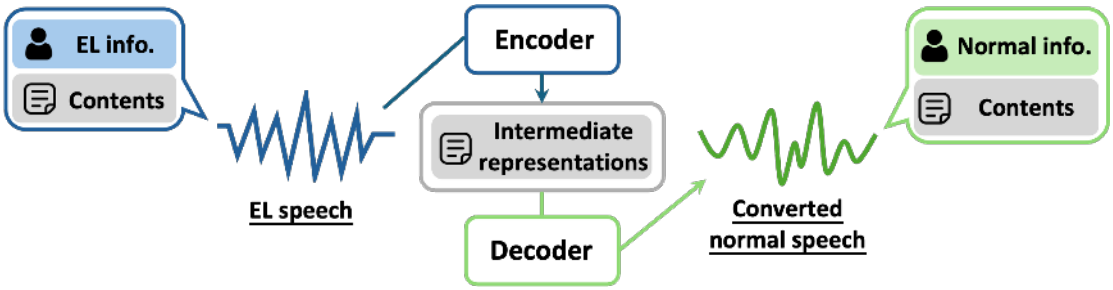


Figure 1.4: Encoder–decoder-based sequence-to-sequence (seq2seq) VC framework for EL2SP.

attempts by Yen *et al.* [40] likewise highlight the advantages of the seq2seq model over the non-seq2seq approach in EL2SP. Furthermore, expanding the vision to the broader speaking-aid fields, including tasks of dysarthric VC [41]–[43], EL speech recognition [11], [44], and EL2SP [40], [45], [46], recent progresses has also largely been attributed to the adoption of seq2seq models.

1.3 Thesis Scope

Building on the above background, this thesis aims to enhance the performance of EL2SP by applying the seq2seq framework. However, seq2seq modeling faces three major challenges, primarily stemming from data mismatch and deficiency discussed in Section 1.2. The following subsections introduce these challenges and the corresponding solutions, most of which have been developed since the late 2010s.

1.3.1 Challenge 1: Data Requirement

Due to the nature of DNNs, in general, seq2seq models require a large amount of high-quality training data to ensure their superiority. Meanwhile, seq2seq VC models still depend on parallel corpora, which makes data requirement more stringent, and hence further increases this challenge. As stated earlier, current datasets available for VC are highly limited, while those used for EL2SP are typically even smaller, only several minutes long, and sometimes semi-parallel, due to strict recording protocols and

the fragile physiological functions of laryngectomees. Such insufficient training data make it difficult for seq2seq models to achieve stable and generalized performance for EL2SP, leading to a decline in the quality of converted speech, such as mispronunciations and repeated or skipped phonemes [47].

One popular approach for addressing limited EL2SP datasize is to leverage transfer learning, which involves the application of pretraining and fine-tuning techniques, collectively referred to as the *pretraining–fine-tuning* paradigm. In this approach, the model is initially pretrained on an extensive, out-of-domain dataset. The pretrained knowledge, encompassing global and high-level feature representations, then serves as a foundation for fine-tuning on a smaller, task-specific dataset, thereby facilitating the overall performance on the target task. In general, pretraining can be divided into two major categories depending on the availability of labeled data: (1) *Supervised* representation learning, and (2) *Unsupervised* representation learning. The latter category also includes self-supervised learning, which relies on a surrogate loss function. In Computer Vision (CV), prevalent ImageNet-based [48] supervised pretraining strategies (e.g., Contrastive Language–Image Pretraining (CLIP) [49] and Supervised Contrastive learning (SupCon) [50]) and some unsupervised strategies (e.g., Momentum Contrast (MoCo) [51] and Bidirectional Encoder representation from image Transformers (BEiT) [52]) serve as robust backbones for downstream fine-tuning tasks such as image classification [53], [54] and generation [55], [56]. Meanwhile, in Natural Language Processing (NLP), the unsupervised manner has demonstrated early successes in enhancing downstream performance, as exemplified by Bidirectional Encoder Representations from Transformers (BERT) [57] and Generative Pretrained Transformer (GPT) families [58]–[61].

Advances in the pretraining approach have also been made in speech processing, drawing inspiration from (1) unsupervised NLP models (e.g., waveform-to-vector (wav2vec)-based ASR [62]), and (2) increasingly accessible and growing labeled databases, enabling the straightforward design of supervised learning systems (e.g., Fast-Speech

[63]) for various downstream tasks. Nonetheless, the effectiveness of unsupervised representation learning for EL2SP remains limited compared to supervised representation learning owing to the intricate nature of data, making it difficult to design a universally effective pretraining objective.

Therefore, this thesis focuses on supervised learning as a promising approach for EL2SP. For developing an EL2SP system using the *pretraining–fine-tuning* paradigm, the typical procedure is to conduct pretraining on a large VC dataset consisting of normal speech pairs, followed by fine-tuning on the small-scale EL2SP dataset. However, this approach faces two problems:

> ***P-1***: Effective pretraining requires a large-scale and parallel VC corpus, which is equally scarce in practice, hindering the development of the pretrained seq2seq VC model.
>
> ***P-2***: Due to inherent differences in feature distributions between EL and normal speech, the huge domain-shifts across the VC and EL2SP datasets would be exhibited, which limits the transferability from pretraining to fine-tuning.

The research question arising from these two problems can be summarized as: **How can we reduce the training data requirement and improve the transfer learning efficacy?**

1.3.2 Challenge 2: Adaptability to Real-world Scenarios

Training complex DNN-based models typically requires extensive datasets that are pure and tailored to the target application. This is particularly true for VC, where consistency in the properties of training data is crucial for achieving effective performance. Aside from the influence of inherent feature distributions, the performance of the resulting models also strongly depends on the quality and structure of the training corpus. Therefore, an ideal training dataset not only needs to be parallel to facilitate

alignment, but also requires the source and target speech pairs to share clean acoustic conditions in order to maintain relatively close feature distributions.

However, obtaining adequately clean training data is difficult in real-world scenarios, which manifests a more practical issue of data deficiency. In such scenarios, the speech signals are often entangled with various background interferences, including noise and reverberation, thereby introducing complex distortions. These distortions cause the distributions of speech-related information, such as linguistic contents and speaker identity, to highly deviate from those in clean speech [64], [65]. Consequently, in real-world scenarios where only distorted data are available for processing, model performance inevitably degrades.

This issue is particularly pronounced in EL2SP. Given the inherent feature disparity between EL and normal speech, building an effective mapping is already challenging. Therefore, current EL2SP developments rely on high-quality datasets recorded in clean studio environments, despite their small scale due to a laborious recording process for laryngectomees. The purpose of this is to eliminate non-speech-related interferences and concentrate solely on the mapping between EL and normal speech. However, this approach introduces a new limitation: The resulting EL2SP systems cannot adapt to acoustically interfered conditions, severely restricting the scope of applicability of EL2SP systems.

Existing works mainly focus on extending the VC framework to improve its robustness for interfered conditions. Considering that any speech technology system relying on clean speech signals can benefit from overlapped speech separation as a front-end processing step, one straightforward approach is to employ Speech Enhancement (SE) [66] as a preprocessing module. This line of work adopts an *enhancement–conversion* pipeline by incorporating extra components such as a denoising module (e.g., Deep Complex Convolution Recurrent Network (DCCRN) [67]) and/or a dereverberation module (e.g., Time-domain audio separation Network (TasNet) [68]), to develop noise- and/or reverberation-robust VC systems [69], [70]. However, one pit-

fall is that introducing additional structures would reduce efficiency in practice compared to single-structure modeling. Furthermore, the performance essentially relies on the prior knowledge of the interferences and the effectiveness of the SE modules. Since SE modules are typically trained on more easily accessible normal speech, their performance is compromised when applied to un-seen EL speech due to the huge differences between EL and normal speech. Additionally, SE processing and transmission inevitably distort speech information [71], which negatively impacts downstream VC. Lastly, most methods introduced above utilize frame-wise VC, which struggles to solve the difficult alignments in EL2SP.

This challenge therefore poses the following research question: **How can we enhance a unified EL2SP system to robustly handle real-world scenarios while using the limited data currently available?**

1.3.3 Challenge 3: Imperfect Internal Representation Learning

As described in Challenge 1 in Subsection 1.3.1, the mismatch between EL and normal speech is identified as a key factor limiting transfer learning. In addition to this, it gives rise a more direct issue within the internal modeling of seq2seq EL2SP, making it inherently more difficult to learn accurate and robust intermediate representations during training.

Specifically, in EL2SP, achieving effective alignment between EL and normal speech is essential for high-quality conversion. As discussed in Subsection 1.2.2, seq2seq VC models establish implicit temporal alignment between the input and output, enabling flexible adjustment of rhythm and prosody. Combined with simple pretraining and fine-tuning strategies, such an approach has demonstrated clear advantages over conventional methods.

However, this paradigm encounters notable performance bottlenecks in EL2SP, since the substantial input–output difference increases the complexity of representation learning. When processing EL speech features, the encoder struggles to extract

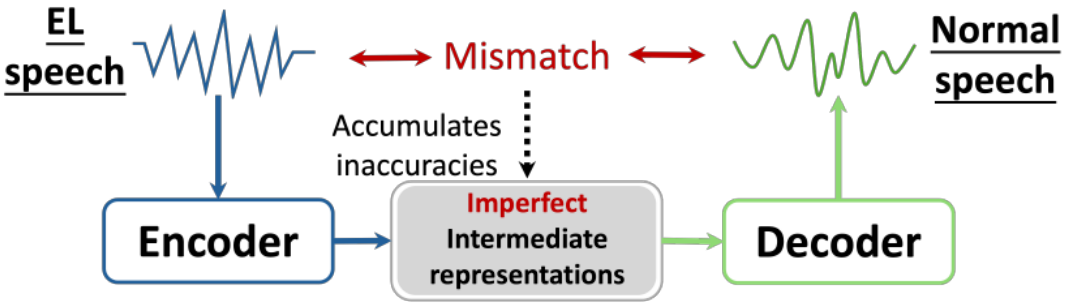


Figure 1.5: Representation mismatch and error accumulation in seq2seq EL2SP training.

pure and stable intermediate representations, inevitably introducing cumulative errors. These errors propagate to the decoder, thereby reducing the accuracy and consistency of the mapping (as illustrated in Figure 1.5). Consequently, there is still much room for the converted speech to be improved before reaching normal-human-level naturalness and intelligibility.

This implies that the representation learning strategy of existing frameworks remains imperfect. Based on this, the research question here is: **How can we reduce the representation mismatch in EL2SP training?**

1.3.4 Potential Line of Solution

Challenge 1 introduced in Subsection 1.3.1 reveals that the typical transfer learning paradigm, namely the direct *pretraining–fine-tuning* approach, is still too naive to be effectively applied in EL2SP. This specific limitation is reflected in two subaspects: *(P-1)* Unavailability of required pretraining data, and *(P-2)* Failing to bridge the huge domain-shift gap under the scarcity of EL2SP data.

With regard to *P-1*, the conventional pretraining strategy must be thoroughly restructured to enable the acquisition of satisfactory initialization parameters even in the absence of large-scale parallel VC corpora, or when no parallel datum is available at all. As noted in Section 1.2, large-scale datasets for training TTS or ASR models are generally more accessible. Therefore, leveraging the datasets from ASR or TTS to assist VC pretraining constitutes a promising solution. For instance, sharing the

speech decoder and attention mechanism from the pretrained TTS model can provide valuable knowledge that is transferable to VC models. This approach opens the new perspective for the pretraining of EL2SP.

Meanwhile, with regard to *P-2*, it can be attributed to the too small amount of task-specific data, which hampers the effective adaptation of pretrained out-of-domain knowledge with significant differences to the target task during fine-tuning. *Data augmentation* can be viewed as an efficient approach. It can increase both the diversity and quantity of the original data by additionally generating synthetic training data, thereby resolving the issues posed by non-parallel, semi-parallel, or limited data. Here, the Synthetic Data (SD) of a greater volume are mainly generated from a well-built speech synthesizer, such as a TTS model. The seminal work by Biadsy *et al.* [72] utilized highly intelligible synthetic speech generated from a high-performance TTS model, to develop an any-to-one VC system for hearing-impaired speech enhancement. A subsequent work by Huang *et al.* [73] employed parallel SD to address the limited non-parallel data available in the VCC2020 dataset [22]. Following this advancement, previous investigation by the author of this thesis [74] further verified the efficacy of SD, which was produced using different portions from the original datasets of normal speakers. Given that *data augmentation* can provide task-specific SD, designing training strategies incorporating large-scale SD can further enhance the transferability of the model from pretraining to fine-tuning. This concept has been validated in some works [11], [44], [45].

Next, for Challenge 2 introduced in Subsection 1.3.2, instead of relying on extra enhancement modules or networks, leveraging more convenient *data augmentation* into seq2seq modeling offers new inspiration for improving the robustness of the model in real-world scenarios. By incorporating data with diverse acoustic and environmnetal attributes into the training process, the model's adaptability to fine-grained real-world conditions can be significantly enhanced. These data variations can be simulated through data augmentation techniques. Similar approaches have been explored [75]–

[78]. Furthermore, as described in Subsection 1.2.2, the encodings produced by seq2seq VC are generally believed to represent pure linguistic information. Given the limited size of the original EL2SP dataset, constructing a pure linguistic representation space with the aid of pretraining would likewise provide a crucial foundation for subsequent adaptive training under real-world scenarios.

Finally, for Challenge 3 introduced in Subsection 1.3.3, a potential solution direction is to enrich the intermediate representation space of the seq2seq model, rather than relying solely on direct mappings between speech features. One promising approach involves incorporating linguistic representations during training. Compared with speech features, linguistic information is inherently independent of speaker-specific characteristics and can provide a clearer, more structured representation of the underlying content. For this purpose, some prior studies have investigated integrating TTS modules into the seq2seq VC framework during the training process. This can be done either to obtain linguistic representations from additional text inputs as auxiliary priors [79], or to align speech and linguistic representations [80], [81] so as to guide the speech encoder toward producing more accurate and robust intermediate representations.

1.4 Thesis Overview

This thesis aims to address the three key challenges based on the potential solution directions outlined above. This primary objective is on developing effective yet streamlined training techniques that can be seamlessly integrated into current seq2seq VC frameworks, while ensuring robustness and efficiency under practical constraints, i.e., data deficiency, environmental interferences, and input–output mismatches. These improvements can highly broaden the applicability of the framework. Figure 1.6 presents the overall scope of this thesis, showing the interconnections between these challenges and the proposed methods. The remainder of this thesis is

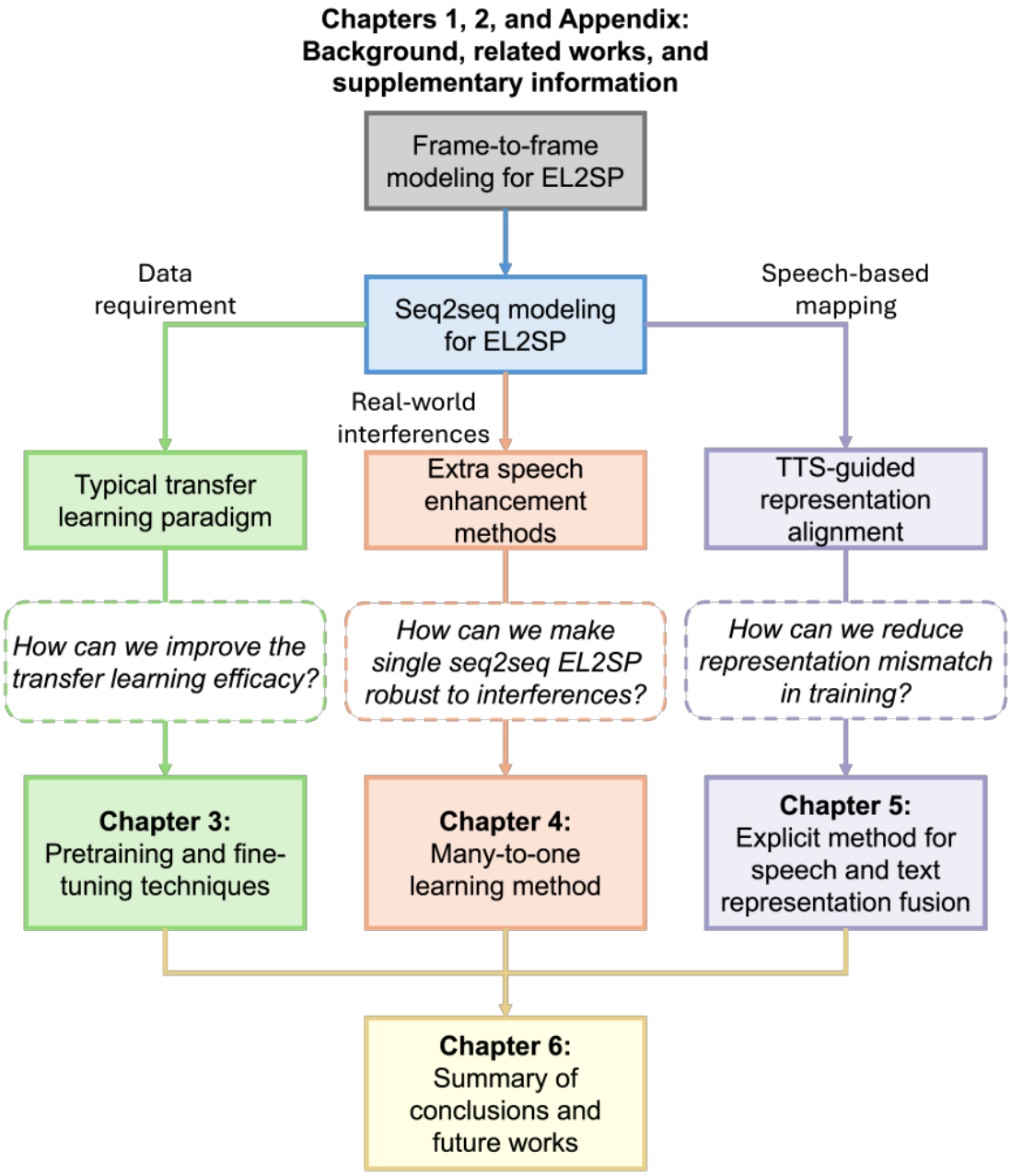


Figure 1.6: Scope and structure of this thesis.

organized as follows:

Chapter 2 is focused on the introduction of the fundamentals of seq2seq VC and reviews the related works underlying the proposed techniques. First, the principles of seq2seq modeling and its primary architectural approach are introduced, which is also adopted in the proposed framework. Next, since the objective of this thesis is to develop novel training techniques to address the aforementioned challenges (Subsec-

tions 1.3.1 to 1.3.3) faced by EL2SP in real-world applications, it is critical to review the essential concepts and strategies involved in related works. This review can provide the necessary background to help readers better understand the formulation of the proposed methods in remaining chapters, as well as the corresponding advancements. Additionally, **Appendix** supplements **Chapter 2** by providing an extended investigation work into the effectiveness of SD for non-parallel seq2seq VC, including additional analyses and empirical observations that further contextualize the related works discussed in the main chapters.

To address the first challenge, **Chapter 3** proposes novel training methods that replace the typical *pretraining–fine-tuning* paradigm. With the power of transfer learning and data augmentation, multistage pretraining and fine-tuning techniques are developed here. Specifically, in light of the potential solution in Section 1.3.4, one essential pretraining stage is introduced by integrating TTS pretraining into VC pretraining, for which the effectiveness has been demonstrated in a supervised method, Voice Transformer Network (VTN) by Huang *et al.* [82]. Furthermore, to strengthen the impact of transfer learning, an encoder adaptation training and a two-stage fine-tuning methods are proposed here, both of which leverage data augmentation. It is worth noting that unlike previous works that rely on high-quality SD, the proposed training techniques require only low-quality SD as training materials. In the pretraining phase, aside from knowledge transfer from a more easily accessible TTS database, encoder adaptation training further minimizes the learning gap of the encoder in comprehending EL speech, facilitating smoother transfer for downstream EL2SP task. Subsequently, the two-stage EL2SP fine-tuning yields a generalized and stable performance. Moreover, by effectively utilizing low-quality SD, the proposed techniques further relax training data demands and enhance practicality. Results on the proposed methods demonstrate substantial performance over the common approach. Meanwhile, comparative analyses confirm progressive performance gains with increasingly deeper system designs.

Next, to address the second challenge, **Chapter 4** presents an interference-robust EL2SP model with a single seq2seq framework. This model directly converts interfered EL speech to clean normal speech without intermediate SE processing. To achieve this, a data-driven training method is designed here, rather than relying on any SE module. Due to the limited size of the original dataset, data augmentation, as indicated in the discussion of Subsection 1.3.4, is employed. Notably, two types of data augmentations are built: (1) Parallel SD for expanding the available data volume, and (2) Real-world condition augmentation in which all EL data are further augmented with background environmental conditions, including noisy, reverberant, and noisy-reverberant conditions, to simulate diverse versions of interfered EL speech. Then, building on a pre-trained VC model, these pseudo-noisy and reverberant EL speech data derived from limited clean data are utilized, together with their normal counterparts, to employ a many-to-one two-stage training scheme, to finalize the model. Based on this framework, different system designs are explored, and their intermediate representations are analyzed to understand their role in filtering the interferences. Furthermore, a systematic comparative study is conducted. For fair comparison, multiple SE-based comparable systems are proposed here, against which the effectiveness and efficiency of the proposed method are demonstrated.

Finally, to address the third challenge, **Chapter 5** extends the mainstream approach —limited to speech-representation-based mapping— by proposing a novel representation learning method that incorporates text representations. The method first integrates a TTS encoder into a seq2seq VC framework to extract text representations, similar to the description in Subsection 1.3.4. However, unlike the previous approach that merely leverages prior knowledge provided by TTS, or indirectly guide the VC encoder output to align with the TTS encoder representations, the proposed method directly fuses text representations with the corresponding EL speech representations to construct richer intermediate representations, which are then explicitly aligned with the target speech features. This fundamentally alleviates the cross-domain represen-

tation mismatch problem in EL2SP. Furthermore, a reconstruction training technique is devised to enable the final EL2SP model to inherit the benefits of this representation enrichment, while maintaining the framework and inference process of the simple seq2seq VC, without increasing complexity. Finally, experimental results show the effectiveness of the proposed method compared with the typical baselines that rely solely on speech-based representations.

In the end, **Chapter 6** provides the summary with the proposed contributions, conclusions, and the implications of this thesis. It also discusses the limitations of the proposed methods and prospects the potential directions for future research areas.

2 | Background and Related Works

This chapter provides the background knowledge that is most relevant to this thesis. The coverage is not intended to be exhaustive, rather, the goal is to quickly familiarize readers with the essential concepts and techniques, facilitating the understanding of the subsequent chapters. Since sequence-to-sequence (seq2seq) Voice Conversion (VC) forms the backbone of this thesis, Section 2.1 introduces its fundamentals, including the model architecture adopted and training losses, followed by the review of transfer learning and data augmentation techniques. Section 2.2 then discusses existing Speech Enhancement (SE) strategies and their applications to interference-robust VC. Section 2.3 briefly reviews works that incorporate linguistic representations through auxiliary modules. In addition, the sections above highlight the differences between these works and the methods proposed in this thesis. Finally, Section 2.4 outlines the commonly adopted evaluation metrics and protocols in this thesis.

2.1 Fundamentals of Sequence-to-sequence (Seq2seq) Voice Conversion (VC)

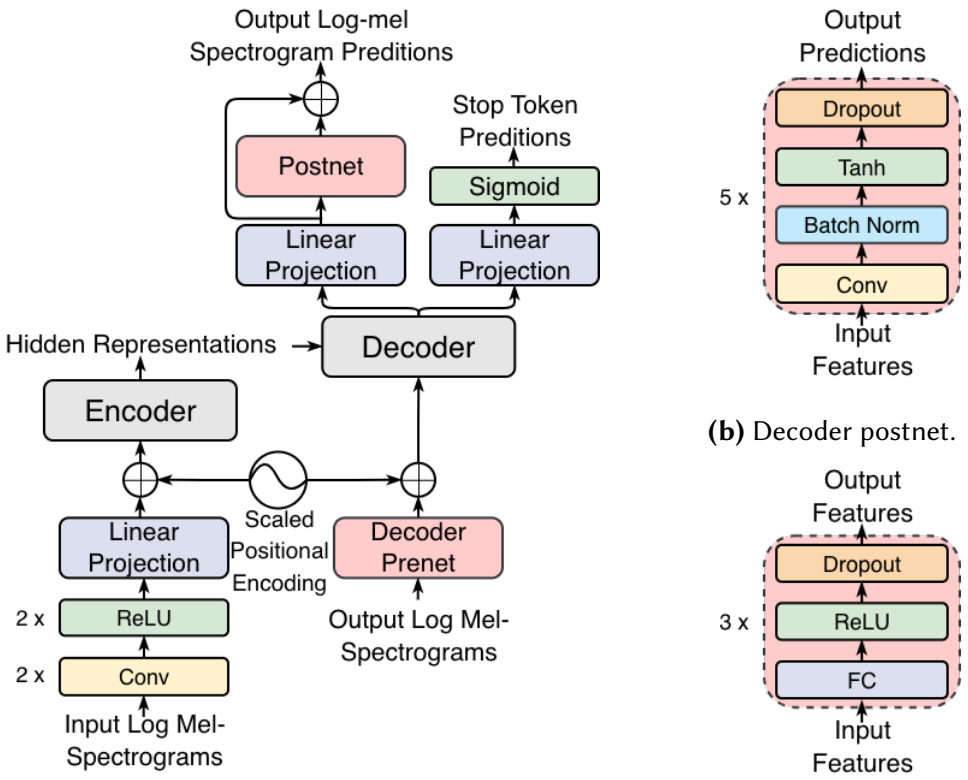
2.1.1 Overview and Motivation

In recent years, seq2seq VC has attracted increasing attention, achieving remarkable performance across a range of VC tasks involving normal speakers [22]. The earliest studies were largely inspired by Recurrent Neural Network (RNN)-based seq2seq

models in neural machine translation [33], [34] and Text-To-Speech (TTS) [83], [84], leading many seq2seq VC models to adopt RNNs [38], [85], such as architectures based on Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs). While effective at capturing long-range dependencies, RNN-based models are hindered by inherently sequential computations, which restrict training efficiency. To mitigate this limitation, Convolutional Neural Networks (CNNs) were introduced in Kameoka *et al.*'s work [86], achieving comparable performance to an RNN-based seq2seq baseline while offering more efficient parallelization. Nonetheless, CNNs rely on the number of different kernels in convolutional layers and still struggle with the same difficulty faced by RNNs in handling long sequences. Against this backdrop, the Transformer architecture [87] has emerged as a compelling alternative. Equipped with a multi-head self-attention mechanism and positional encodings, Transformers not only accelerate training through fully parallelizable operations but also boost the perception for global contextual dependencies. These advantages have been validated in pioneering works on both TTS [88], [89] and VC [82], where Transformer-based seq2seq models have demonstrated strong performance.

Leveraging the successes of a seq2seq model to normal-to-normal VC, researchers have also extended seq2seq VC models to the speaking-aid field, aiming to convert disordered or impaired speech [40], [42], [72], [90], [91]. However, despite these efforts, the contributions of seq2seq VC to enhance EL speech —one of the most representative types of impaired speech— remain limited, especially for constructing efficient intermediate representations when only small in-domain datasets are available.

Given these developments, and considering the need for efficient modeling in EL-speech-to-normal-SPeech conversion (EL2SP), this thesis adopts a Transformer-based seq2seq framework as the primary architecture, with the goal of addressing the unique challenges posed by EL2SP. The following subsections first provide an overview of the seq2seq VC model (Subsection 2.1.2), then describe the Transformer-based components in detail (Subsection 2.1.3), and finally introduce the training objectives (Sub-



(a) Core architecture and the extensions.

(b) Decoder postnet.

(c) Decoder prenet.

Figure 2.1: Illustration of the overall seq2seq VC framework.

section 2.1.4).

2.1.2 Overall Framework of Seq2seq VC

As depicted in Figure 2.1(a), the seq2seq VC framework is in general, organized into two primary components: Encoder and decoder, each composed of a stack of several identical building blocks. This structural design enables the model to process and transform variable-length speech sequences, a key requirement for VC tasks where the temporal alignment between source and target signals is inherently inconsistent. In practice, seq2seq VC models operate on acoustic feature representations rather than raw waveforms. A common choice is the log mel-spectrogram, which provides a compact yet perceptually relevant representation of speech. In line with this convention, this thesis also adopts log mel-spectrogram, where the source and target utterances

are both transformed into sequences of such feature frames, as the input and output.

Formally, the training objective of the seq2seq VC model is to learn a mapping between source and target speech sequences that are not explicitly aligned at the frame level. Let the source and target sequences be denoted as:

$$X = \left(\mathbf{x}_n \in \mathbb{R}^D \,\middle|\, n = 1, \ldots, N\right), \quad Y = \left(\mathbf{y}_m \in \mathbb{R}^D \,\middle|\, m = 1, \ldots, M\right), \tag{2.1}$$

where N and M represent the length of each sequence, respectively, and D denotes the acoustic feature dimension. In most cases, $N \neq M$, which reflects the natural differences in speaking rate, rhythm, and prosody.

The input speech feature sequence X is first fed into the encoder. The encoder is then responsible for transforming X into a sequence of context vectors of equal length, also referred to as *hidden representations* as:

$$H = \left(\mathbf{h}_n \in \mathbb{R}^D \,\middle|\, n = 1, \ldots, N\right) = \text{Encoder}(X), \tag{2.2}$$

which capture contextual and linguistic information, serving as a bridge between the source features and the generation process carried out by the decoder.

Afterwards, the decoder, in turn, operates in an autoregressive manner to build contextually relevant inference and sequentially generate the target speech sequence. At each time step t, the decoder predicts the current outputs $\mathbf{y}_t$ by conditioning not only on the encoded hidden representations H but also on all previously generated outputs $(\mathbf{y}_1, \ldots, \mathbf{y}_{t-1})$. This process can be formally expressed as:

$$\mathbf{y}_t = \text{Decoder}(H, (1, ..., \mathbf{y}_{t-1})). \tag{2.3}$$

This recursive dependency allows the model to capture temporal dynamics and maintain consistency across successive frames of the generated speech sequence.

Beyond the basic encoder–decoder architecture, seq2seq VC models often integrate several auxiliary components to enhance performance and stabilize training, a

practice largely inspired by advances in modern seq2seq TTS frameworks [83], [84]. These additional modules are particularly important in EL2SP, where the complexity of acoustic mapping and the variability of input signals increase the difficulty of effective learning.

First, convolutional layers are commonly employed prior to the encoder in order to preprocess the acoustic features [92]. In this thesis, two convolutional blocks with 3×3 kernels and 2×2 strides are applied, each followed by a Rectified Linear Unit (ReLU) activation. This setting downsamples the input log mel-spectrograms along both time and frequency axes, reducing them to one quarter of the original size. The downsampling not only streamlines the computations in subsequent layers but also emphasizes salient phoneme-level information [85], which in turn facilitates the learning of attention alignments. A subsequent linear projection adjusts the dimensionality of the convolutional outputs to match that of the positional encoding, after which the features enriched with positional information are passed into the encoder.

On the decoder side, a *prenet* is introduced as an essential information bottleneck for the autoregressive process. Specifically, it consists of two fully connected layers with ReLU activations and dropout layers, and it operates on the predictions from the previous time step. By constraining the flow of information, the prenet can improve the robustness of attention learning and facilitate the formation of stable diagonal alignments. Following the decoder, two parallel linear projections are applied to separately process the decoder outputs. One branch, combined with a sigmoid activation function, predicts a stop token to determine the end-of-speech generation, thereby enabling an adaptive inference process rather than relying on a fixed-length output. The other branch directs the decoder outputs to the *postnet*. The postnet is composed of five convolutional layers with 5×5 kernels, each followed by batch normalization, a hyperbolic tangent (tanh) activation, and a dropout layer. This deeper structure enables the network to capture more profound contextual dependencies from the projected outputs. Note that the postnet outputs are then added to the linearly projected

decoder outputs, yielding the final refined spectrogram predictions. The blocks of the decoder prenet and postnet are illustrated in Figures 2.1(b) and (c), respectively.

Furthermore, employing a reduction factor can improve the efficiency of seq2seq training. By allowing the decoder to predict multiple consecutive frames (as determined by the reduction factor) at each step instead of a single frame, this approach accelerates convergence, reduces computational cost, and leverages the correlations that naturally exist among neighboring speech frames, thereby promoting more effective interaction modeling with the hidden representations.

2.1.3 Transformer-based Architecture

The seq2seq VC model in this thesis adopts a structure similar to the Voice Transformer Network (VTN) [82], which is a representative application of the Transformer [87] backbone to voice conversion. To provide a clear understanding of the adopted Transformer architecture, the main interconnected components, organized according to the data flow illustrated in Figure 2.2, are described below.

Scaled positional encoding: Before feeding feature sequences to the encoder or decoder, positional information must be injected for each element, which is crucial for Transformers. Unlike RNNs, the original Transformer processes inputs in parallel and therefore cannot inherently perceive the position or order of elements in the feature sequences. A common remedy is the sinusoidal Positional Encoding (PE) function [87], which provides an inductive bias for position. In the proposed seq2seq model, Scaled Positional Encoding (SPE) [88] is followed by further introducing a trainable scalar weight α to the PE function, so that the encodings can better adapt to the input scales of both the encoder and decoder. The encoding of time step t in a sequence is computed as:

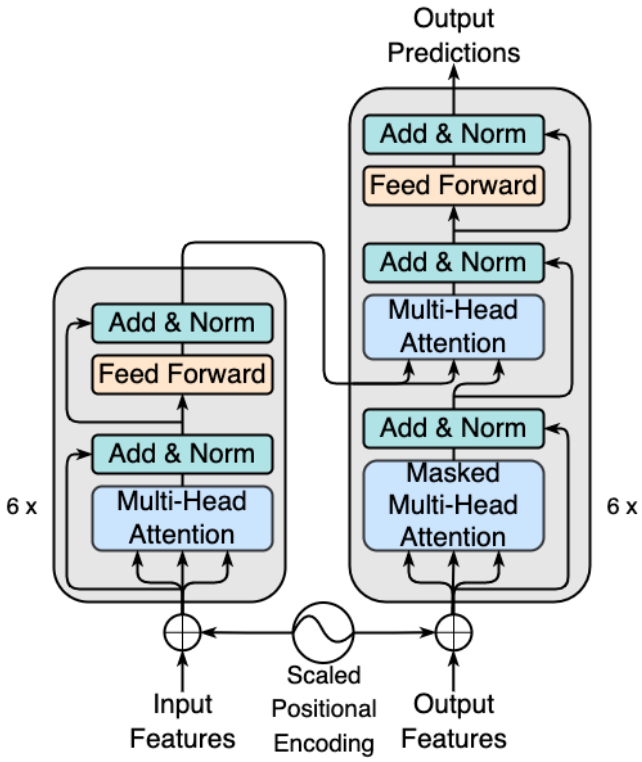


Figure 2.2: Illustration of the Transformer architecture.

$$\begin{aligned} \mathrm{SPE}(t, 2i) &= \alpha \sin\left(\frac{t}{10000^{\frac{2i}{d_{\text{model}}}}}\right), \\ \mathrm{SPE}(t, 2i+1) &= \alpha \cos\left(\frac{t}{10000^{\frac{2i}{d_{\text{model}}}}}\right), \end{aligned} \tag{2.4}$$

where d_{model} denotes the feature dimension of the model, and $i = 0, 1, \ldots, \lfloor d_{\text{model}}/2 \rfloor - 1$ indexes the sinusoidal pair, corresponding to dimensions $2i$ and $2i+1$.

Transformer-based encoder: The d_{model}-dimensional feature sequences enhanced with SPE serve as the initial input to the encoder, which consists of six identical layers stacked together. Each layer has two sublayers, namely *Multi-Head Attention* (MHA) and *position-wise Feed-Forward Network* (FFN), either of which is wrapped with *residual connection* and *layer normalization* [93], as described below:

- *MHA sublayer*: The input matrices: query Q, key K, and value V are transformed by three trainable weight matrices $W^{(Q)}$, $W^{(K)}$, and $W^{(V)}$, respectively. Then, these transformed matrices are split into h *self-attention* heads to perform

independent and parallel attention learning. The output values from each head, referred to as head_i, are concatenated and once again projected by another parameter matrix $W^{(O)}$, to generate the final MHA outputs. The attention function adopts the scaled dot-product attention, which was demonstrated to be effective by Vaswani *et al.* [87] as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.5}$$

where d_k is the dimension of K. The MHA function including head_i is formulated as:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^{(O)}, \tag{2.6}$$

$$\text{where head}_i = \text{Attention}(QW_i^{(Q)}, KW_i^{(K)}, VW_i^{(V)}). \tag{2.7}$$

Note that the dimensions of Q, K, V are all equal to d_{model}, and the dimension of head_i is d_{model}/h. With the above mechanism, different aspects of the input sequences can be jointly attended to at the same time.

- *FFN sublayer*: Each FFN layer is composed of two linear transformations separated by a ReLU activation layer. By using different parameters from layer to layer, this FFN structure can be applied identically and independently to each position of the sequence to enhance the model's representational capacity. This process is defined as:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{2.8}$$

where x represents the input sequence. It undergoes two linear transformations with learnable weight parameters W_1 and W_2, and corresponding bias terms b_1 and b_2, respectively, with the ReLU activation in between.

- *Residual connection and layer normalization*: The residual connection followed

by layer normalization is applied around each of the aforementioned two sublayers. Specifically, the output is obtained by normalizing the combination of the input and output of each sublayer. The input vector X and the sublayer function Sublayer(X) are used to determine the normalized output of each sublayer, which can be stated as:

$$\text{LayerNorm}(\text{Sublayer}(X) + X). \tag{2.9}$$

Transformer-based decoder: The decoder, likewise, consists of six identical layers, each containing the same two sublayers (MHA and FFN sublayers) as in the encoder layer, while an additional *masked* MHA sublayer is inserted. As in the encoder, each sublayer is wrapped with a *residual connection* and then passed through a *layer normalization*. Exclusive decoder sublayers are described below:

- *Masked MHA sublayer*: This sublayer enforces causality by utilizing the masking mechanism, allowing the attention only to the preceding and current positions in the output sequence, and excluding the future ones. Thus, the autoregressive generation property of the seq2seq model is preserved.

- *Encoder–decoder MHA sublayer*: In contrast to the MHA with the *self-attention* manner in the encoder, the second MHA sublayer in the decoder jointly processes the encoder output H as the memory keys and values with the queries derived from the previous masked MHA sublayer, thereby building *encoder–decoder attention* for the Transformer. This design allows the decoder to integrate the encoded information of the input sequence from multiple perspectives when generating the output.

2.1.4 Training Objectives

This thesis primarily employs three types of loss functions to optimize the seq2seq network. The L1 loss (l_1) is used for prediction of continuous feature sequences, while the weighted Binary Cross-Entropy (BCE) loss (l_{BCE}) is used for stop token prediction. Furthermore, an additional issue identified by Liu *et al.* [94] is addressed here. Specifically, the first two loss functions cannot reliably enforce diagonal attention patterns across all attention heads in the Transformer-based model, which may cause a significantly slower convergence. To mitigate this problem, a Guided Attention loss (l_{GA}) [38], [94]–[96] is applied here to the partial encoder and decoder layers, which encourages the attention matrix for a given input–output sequence pair to align more closely with the diagonal structure defined by the corresponding guide matrix. Let $\hat{I}$ and I denote the predicted and target feature sequences, respectively. L1 loss can be denoted as:

$$l_1(I, \hat{I}) = |I - \hat{I}|. \tag{2.10}$$

Let $\hat{P}$ and P denote the stop token probability and target binary label, respectively. BCE loss can be denoted as:

$$l_{\text{BCE}}(P, \hat{P}) = -W \left(P \log(\hat{P}) + (1 - P) \log(1 - \hat{P}) \right). \tag{2.11}$$

Let the attention matrix $A \in \mathbb{R}^{N \times M}$ correspond to the pair of source and target sequences X and Y, where N and M denote the input and output lengths, respectively. To encourage a diagonal alignment, a guide matrix $G \in \mathbb{R}^{N \times M}$ is defined, whose (i, j)-th element is given as:

$$g_{i,j} = 1 - \exp\left(-\frac{\left(\frac{i}{N} - \frac{j}{M} \right)^2}{2\sigma_g^2} \right), \tag{2.12}$$

where σ_g controls the width of the diagonal band, determining how strictly the attention is encouraged to stay near the diagonal. Guided Attention loss is then formulated

as:

$$l_{\mathrm{GA}} = \lambda_{\mathrm{GA}} \|G \odot A\|_1, \tag{2.13}$$

where $\odot$ denotes element-wise multiplication and λ_{GA} is the weighting coefficient.

The total loss function is formulated as:

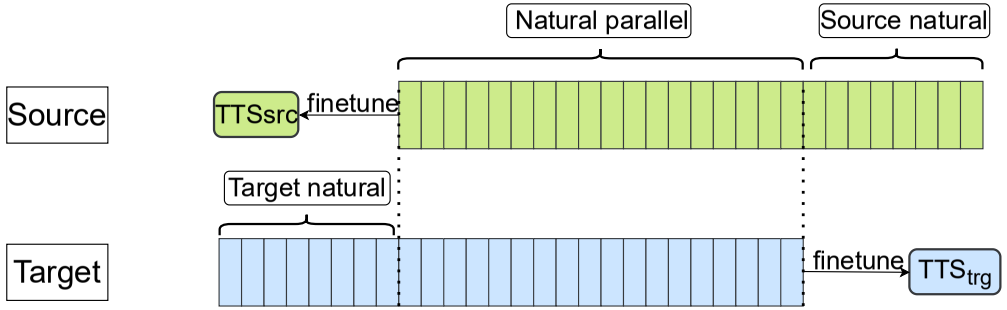
$$l_{\mathrm{total}} = l_{\mathrm{seq}} + l_{\mathrm{token}} + l_{\mathrm{GA}}, \tag{2.14}$$

where l_{seq} represents the feature sequence loss derived by summing the L1 losses l_1 of the projected outputs and of postnet outputs, and the stop token prediction loss l_{token} is computed using the BCE losses l_{BCE} of the linearly projected tokens and of the sigmoid activated tokens.

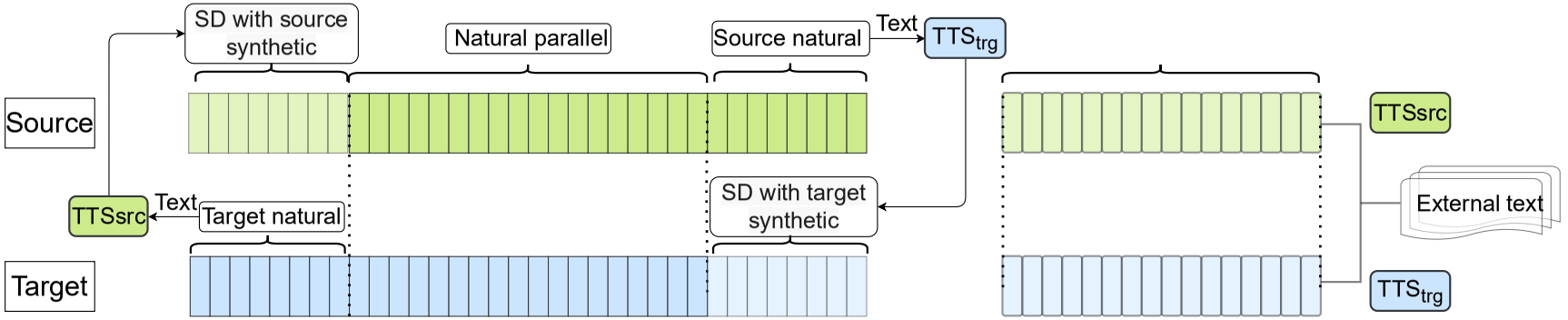
2.1.5 Data Augmentation in VC

As mentioned in Subsection 1.3.4, data augmentation is a common approach to address the deficiency of initial training data in VC. Aside from leveraging TTS models to generate Synthetic Data (SD) with greater size and diversity for VC training, conversely, VC itself can also serve as an augmentation technique to assist other speech-processing tasks in low-resource scenarios [97], [98].

In this thesis, the training techniques proposed across Chapters 3 to 5 all incorporate data augmentation as a component of their system designs. Here, the most relevant works are introduced. Studies by Huang *et al.* [73] and the author of this thesis [74] demonstrated that incorporating TTS-generated SD as an extension of the original non-parallel or semi-parallel data settings can facilitate the training of a seq2seq VC model. Figure 2.3 shows the basic concept, where parallel SD is synthesized under an originally semi-parallel dataset. First, separate TTS models for the source and target speakers are developed by fine-tuning a pretrained TTS model using their respective natural speech data from the semi-parallel dataset. Then, source SD and target SD are generated by using the trained TTS models. Consequently, in addition to the original



(a) Text-To-Speech (TTS) training process using the semiparallel dataset.



(b) Parallel SD generation process using source SD, target SD, and external parallel SD.

Figure 2.3: Two steps of the generation process of a parallel Synthetic Data (SD) from a semiparallel dataset, previously proposed by the author of this thesis [74].

natural parallel data in the semi-parallel dataset, two types of training pairs containing SD are also generated, which are ⟨source synthetic, target natural⟩ and ⟨target synthetic, source natural⟩. Moreover, it is possible to additionally introduce external text to produce external parallel SD to further enhance the training process on the originally imbalanced datasets. The investigation of TTS-generation-based data augmentation under normal speech sets is presented in detail in Appendix.

Furthermore, Chen *et al.* [99] introduced ParaGen, a data augmentation technique for non-parallel VC. With extra speaker disentangling modules, this method eliminates unnecessary acoustic components to generate augmented data that preserve the target speaker's identity and the source speaker's speaking style, achieving a direct frame-to-frame VC. However, this approach requires a sophisticated architecture with rigorous hyperparameter tuning to guarantee clean and high-quality augmented data, otherwise, the performance will degrade. Biadsy *et al.* [72] presented an any-to-one VC system named Parrotron to normalize an arbitrary source speech into a single, canonical target speech. It is trained on a synthetic corpus generated by a powerful WaveNet-based TTS system [100], and used to convert the highly atypical speech of a deaf speaker into a more intelligible and natural speech. However, building the TTS model requires a huge amount of training data to ensure the quality of output samples. Closely related to the EL2SP field, Yang *et al.* [91] proposed an Automatic Speech Recognition (ASR)-style Bottleneck Feature (BF) extractor to obtain accurate EL and normal speech features for training an EL2SP system. To enhance the training, the WORLD vocoder [29] was utilized to flatten the F0 contour of a normal speech dataset to synthesize a simulated EL speech set for data augmentation. However, the effectiveness of the system relies on data sufficiency, as the extra BF extractor is trained on a large EL natural speech dataset. Moreover, accurate configurations are also required to modify the normal speech.

The training techniques proposed in this thesis are different from the aforementioned works. They share a unified architectural principle: The performance of the

resulting systems are built solely on a single seq2seq model and is independent of external modules. Moreover, except for using a normal corpus as a base *a priori* from scratch, the proposed methods primarily leverage imperfect SD to enhance training strategies, particularly in pretraining and fine-tuning, without requiring high-quality SD. This design contributes to improved robustness and transferability in addressing the three challenges outlined in Subsections 1.3.1 through 1.3.3.

2.1.6 Transfer Learning for VC

Transfer learning requires a sizable pretraining dataset to generalize well, but collecting a considerable amount of parallel VC corpus is difficult. To accomplish pretraining, one direction is to utilize the TTS architecture, as the pure linguistic intermediate representations and high-quality speech reconstruction mechanism it provides are essential for VC. Motivated by a large and easy-to-access TTS corpora, Park *et al.* [81] performed a multi-speaker TTS pretraining to build a transcription-guided speech encoder, thus improving the accuracy of encoded features for any-to-many VC. Similarly, Wang *et al.* [101] proposed an end-to-end dysarthric speech reconstruction framework, where a pretrained TTS text encoder guides a speech encoder via cross-modal knowledge distillation to extract robust linguistic representations from impaired speech. Another prevalent concept is sharing the attention and decoder of TTS, which was reported by Zhang *et al.* [79], [102], and Huang *et al.* [82]. Additionally, a joint training method for VC and TTS is proposed by Zhang *et al.* [79], providing the hybrid benefits to the VC task. A new research perspective by Luong and Yamagishi [80] suggests that a non-parallel VC can be bootstrapped by the pretraining of the speaker-adaptive TTS and an unsupervised acoustic encoder. On the other hand, the well-pretrained ASR can effectively disentangle linguistic contents from the speaker-specific information of speech, making it another choice for developing VC. Leveraging extra modules, such as the pretrained Phonetic PosteriorGrams (PPGs) or BF extractors derived from the acoustic model of ASR, can aid in the VC systems [91], [103]–[106]. Furthermore,

pretraining the ASR encoder to initialize VC, or cascading ASR with the TTS has been demonstrated to be effective [36], [107]. Aside from these two main directions, Yang *et al.* [91] conducted VC pretraining using parallel augmented data.

This thesis incorporates transfer learning as a foundational strategy. In contrast to prior works, the methods developed in this thesis neither require an extensive, high-quality VC corpus nor rely on complex model structures. Instead, they adopt a straightforward TTS-oriented perspective, leveraging imperfect but more easily accessible SD in combination with typical TTS pretraining. Furthermore, to address the practical challenges in EL2SP, multistage training strategies are designed tailored to different task-specific domains, data constraints, and learning objectives. This structured adaptation process enhances model robustness and generalizability across real-world scenarios, and has not been explored in previous studies.

2.2 Interference-robust VC

Since this thesis also aims to tackle the robustness of EL2SP systems under real-world conditions with interference, it is necessary to review the relevant works on interference-robust VC. In what follows, three widely studied categories are detailed: (1) Statistical methods, (2) Speech Enhancement (SE) methods, and (3) Representation learning methods. Whereas the seminal works of the latter two were originally applied in the context of TTS [66], [108], [109], the discussion here mainly focuses on the most relevant literature in VC. Table 2.1 summarizes the three categories of interference-robust VC methods.

2.2.1 Interference-robust VC with Statistical Methods

Leveraging the sparse representations based on the Non-negative Matrix Factorization (NMF) function [110] is a common statistical approach for developing interference-robust VC. Takashima *et al.* [111] proposed an exemplar-based VC, wherein NMF

Table 2.1: Summary of interference-robust VC methods.

Method	Statistical-based	SE-based	Representation learning-based
Related subsection	Subsection 2.2.1	Subsection 2.2.2	Subsection 2.2.3
Concept	Exemplar-based VC using NMF	Interference suppression via SE	Learning interference-invariant representations
Scenario	Noisy	Noisy and/or reverberant	Noisy and/or reverberant
Limitation	High computational cost; suboptimal performance	Potential data distortion; dependent on SE quality; extensive speech training data	Extra non-speech training materials

decomposes the spectral features of the acoustic signals into a linear combination of sparsely represented exemplars and their corresponding weight vectors. During inference, the noisy speech, comprising both noise and speech exemplars, is converted into the clean target speech by using target exemplars and the weights of source exemplars. However, besides the issues of its own frame-wise architecture, NMF is a computationally intensive algorithm that requires rigorous parameters and high training costs to obtain accurate sparse representations. Although Aihara *et al.* [112], [113] endeavored to reduce the reliance on parallel data and improve training efficiency, their methods still cannot outperform other conventional VC models.

2.2.2 Interference-robust VC with Speech Enhancement Methods

SE methods, which involve connecting external SE modules or processing stages, are the mainstream approach to address speech interferences. Miao *et al.* [114] realized a noise-robust VC but it requires complex dual noise-filtering strategies for preprocessing and postprocessing. In preprocessing, low-pass filtering is employed to eliminate noise in inputs. In postprocessing, Mel-CEPstral coefficients (MCEPs) undergo statistical filtering to reduce noise in converted coefficients. Furthermore, the input MCEPs are extended and only the sub-band cepstrum is converted to mitigate interferences in high-frequency components. This method, although somewhat effective, relies heavily

on intricate filtering techniques to handle noise throughout the VC process. In contrast, Xie *et al.* [69] developed a noise-controllable VC framework based on a different approach. A pretrained denoising model is firstly utilized to separate noisy speech into noise and speech signals. Subsequently, the downstream Vector Quantized-Variational AutoEncoder (VQ-VAE)-based VC [115] is trained on the denoised speech, eliminating the need for clean training data. During inference, the separated noise can be selectively superimposed onto the converted speech on the basis of specific scenarios. However, since the quality of the denoised speech used for VC training is inferior to that of clean speech, the VC performance is degraded. To reduce such impact, Xie *et al.* [116] and Xie and Toda [117] successively proposed several improvements, such as using the separated noise as a VC condition to directly model the noisy speech and implementing diverse data augmentations, all of which entail increased architectural complexity and training costs.

To sum up, because SE and downstream VC have independent architectures and require different features, additional feature transformations are often necessary in the aforementioned works, thus causing feature distortions. Although Chan *et al.* [118], focusing on noisy condition, designed a lightweight SE module to achieve joint training with VC components incorporating Generative Adversarial Networks (GANs) [119] and various loss functions, such multi-component models still require accurate configurations and balanced loss weights. In addition, under more complex interfered conditions where noise and reverberation coexist, a more comprehensive SE module should be designed to handle a broader range of distortions [70]. On the other hand, the authors of some VC studies [69], [70], [116], [117] affirmed that they did not rely on clean training data because of the integration of the fixed SE module, but the SE pretraining still requires extensive clean speech data to ensure its performance. However, most real-world datasets like AMI [120] and CHIME5 [121], do not provide clean counterparts for the interfered speech recordings.

2.2.3 Interference-robust VC with Representation Learning Methods

The primary objective of representation learning methods is to enhance deep perceptual insights into interfered data. Autoencoder-style denoising models [122], [123] provided some early inspirations in this direction. Afterwards, an attempt to address the noisy condition using GAN-based domain adversarial training was proposed in one-shot VC [124], where two groups of gradient reversal layers and domain classifiers were assigned to the speaker and content encoders, respectively [125]. Therefore, the training objective included not only the reconstruction loss but also domain classification losses. By learning encoded features that are invariant to noise, the framework can handle unseen noise types. Despite this, the need for clearly labeled noisy/clean conditions and sufficient training data remains as potential issues. On the other hand, a few number of studies considered overcoming the reverberation. Huang *et al.* [75] explored general interference-robust VC that combines adversarial and denoising training to tackle noise and reverberation. To achieve effective adversarial training, the embedding attack [126] is additionally used to generate adversarial samples, which are distributed to each mini-batch together with other types of data augmentation. Although the work demonstrated preliminary robustness, some of its case studies revealed adverse impacts. Mottini *et al.* [127] proposed a VC framework that can overcome noisy-reverberant condition. However, besides acoustic signals, this framework requires transcriptions, which are consumed by extra phonetic and ASR encoders for providing textual information, to enhance the efficiency of representation learning. In a similar study, Choi *et al.* [128] proposed a reverberation-robust VC consisting of a VC module and a reverberation time (T60) estimator, which introduces essential T60 information for realizing controllable reverberation.

Altogether, representation learning methods have mainly achieved a single-shot training process without significant data distortion compared with SE methods. How-

ever, most of them still depend on sophisticated frameworks involving either GANs or multiple components, which necessitates supplementary training data and labels. Moreover, the application of these studies to real-world EL2SP is extremely rare.

The method proposed in this thesis, has some similarities with representation learning methods, but unlike all the aforementioned methods for addressing real-world scenarios, the proposed method is more convenient and efficient in various aspects, as follows:

- Data processing. Since the proposed method requires speech features exclusively and simultaneously accomplishes both SE and VC, a unified preprocessing for speech features is carried out, eliminating the need for additional intermediate data processing and analysis.

- Data augmentation. Data augmentation is a straightforward approach to increase the diversity of initial datasets. As mentioned in Subsection 2.1.5, for clean EL2SP, Yang *et al.* [91] utilized synthesized EL speech to enlarge the volume of training data. Similarly, Xie *et al.* [69] and Huang *et al.* [126] developed interference-robust VC by simulating different interfered conditions from clean speech sets. However, these studies encountered a common limitation: Yang *et al.* presumed the availability of ample EL speech for pretraining crucial components, while Xie *et al.* and Huang *et al.* relied on a large amount of high-quality clean data. In contrast, data augmentation in the method proposed in this thesis is a significant step towards more flexible and practical interference-robust EL2SP systems. Two types of data augmentation are considered, i.e., (1) increasing the amount of essential speech data according to imperfect parallel SD generated by fine-tuned TTS models, then (2) simulating interfered EL data by adding diverse real-world acoustic scenarios into the expanded EL data (both the original and imperfectly synthetic EL data) built.

- Framework. The proposed method, only focusing on seq2seq VC, offers more

promising and simplified applications than those requiring complex frameworks. Furthermore, compared with the mainstream seq2seq VC works relying on RNNs [38] or CNNs [86], the proposed method is structurally more suitable for handling real-world scenarios by leveraging the strengths of the Transformer [87]. This not only accelerates training efficiency but also boosts a deep understanding and modeling of different types of data, as evidenced by its success across multiple datasets in Large Language Models (LLMs) [60]. Moreover, combined with data augmentation covering various real-world conditions, the seq2seq-based training techniques designed in this thesis are simple yet effective, as will be elaborated in Chapter 4.

2.3 Representation Learning Leveraging Auxiliary Information

To achieve better mapping performance, especially in the EL2SP task where the source and target speech exhibit large acoustic discrepancies and the available training data are limited, incorporating additional training information to assist in representation learning is a potential direction. As pointed out in Subsection 1.3.4, text is a natural choice, since it inherently removes acoustic properties such as prosody, speaking rate, and pitch, while retaining pure, speaker-independent token-level linguistic information, which is more conducive to model recognition. Some works [79]–[81], [101] mentioned in Subsection 2.1.6 provide a fundamental idea, namely, introducing TTS modules to indirectly exploit text representations, including dual-attention sharing between TTS and VC and text-based alignment supervision. However, such an approach typically relies on well-trained TTS modules on large-scale multi-speaker corpora, as well as complex frameworks and training algorithms. Even so, it is still difficult to ensure (1) the extraction of speaker-independent features during conversion [80], [81], [101], and (2) the balanced utilization of multi-source representations

[79]. These issues limit the overall performance's upper bound and imply that such methods are difficult to be directly applied to EL2SP using EL speech as input. The essential reason here is that text is not directly involved in the modeling as an object that is equally integrated with speech, but rather introduced as a prior condition, an alignment reference, or a multi-source constraint. This indirectness prevents the full exploitation of text advantages, leaving representation learning and conversion performance constrained under conditions of high acoustic mismatch and limited data.

In contrast, explicitly fusing additional modalities (e.g., texts or images) into speech representations provides a promising direction and has achieved progress in subtasks of speech emotion recognition [129]–[131], TTS [132], [133], and VC [134]–[136]. Here, the relevant literature on the latter two are introduced. Chen *et al.* [132] proposed StyleFusion TTS, which incorporates text, reference audio, and speaker timbre simultaneously into a zero-shot synthesis framework. Through a hierarchical fusion module, it achieves highly natural multimodal speech synthesis. Furthermore, Guan *et al.* [133] unified text, speech, and face prompts in a shared style space through an aligned multi-modal prompt encoder, achieving flexible and expressive TTS with style controllability across modalities. On the other hand, in the VC domain, Kameoka *et al.* [134] proposed crossmodal VC to pioneer the joint modeling of speech and facial images, realizing cross-modal conversion between auditory and visual domains. Additionally, Niu *et al.* [135] combined text and audio prompts, aligning style embeddings in the latent space with contrastive learning to enhance flexibility and efficiency in speech style conversion. As another example, Li *et al.* [136] introduced rhythm as an independent modality in speech-to-singing conversion and explicitly fuses speech content with melody information via cross-modal alignment, thereby improving both quality and naturalness. Overall, these studies, focusing on style modeling in TTS and VC, are mainly built upon prompt-based modality guidance, emphasizing the use of text or audio prompts to control the generation process.

Drawn inspiration from the aforementioned studies, this thesis is the first work

in EL2SP to fuse text and EL speech representations within a seq2seq VC framework, leveraging a straightforward TTS encoder.

2.4 Evaluation Metrics and Protocols for the Proposed Methods

In order to evaluate the performance of the proposed EL2SP methods, this thesis employs both objective and subjective evaluation metrics. Objective metrics are widely used in speech processing fields such as VC and TTS, to provide quantitative and repeatable indicators of system performance, which are particularly useful during model development and for large-scale comparisons. Subjective evaluation tests, on the other hand, directly reflect human perception, which remains the ultimate standard for assessing speech quality and intelligibility. In this thesis, objective and subjective evaluations are considered complementary; the former offers efficiency and reproducibility, while the latter captures perceptual aspects that are difficult to fully reflect with current objective measures. In what follows, the objective evaluation metrics used in this thesis are first introduced, followed by the subjective evaluation protocol.

2.4.1 Objective Evaluation Metrics

This thesis adopts the following objective evaluation metrics to assess essential aspects, including spectral quality, intelligibility, and pitch accuracy.

- Mel-Cepstrum Distortion [137] (MCD, in dB): This metric is commonly used to measure the spectral envelope distortion by comparing the ground-truth samples with the converted samples in VC on the basis of the L2-norm concept. It is defined as:

$$\text{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{D}\left(c_i^{(\text{ref})} - c_i^{(\text{gen})}\right)^2}, \tag{2.15}$$

where D denotes the feature dimension of the Mel-Cepstral Coefficients (MCCs), and $c_i^{(\text{gen})}$ and $c_i^{(\text{ref})}$ represent the i-th dimensional coefficients of the generated and target original MCCs, respectively. A low MCD value signifies a smaller distortion, indicating a higher general quality for the converted speech.

- Character Error Rate (CER, in %): Since the methods proposed in this thesis are applied to the Japanese language, CER is applied to evaluate the intelligibility accuracy and character consistency of the generated speech from a character-level perspective. A low CER value indicates higher intelligibility accuracy.

- Scale-Invariant Signal-to-Distortion Ratio (SI-SDR, in dB) [138]: This metric is used to assess the quality of audio signals independent of their scale. It measures the energy ratio between the original and the distortion components of the processed signal after scale alignment. A higher SI-SDR value indicates smaller distortion and higher signal fidelity.

- Short-Time Objective Intelligibility (STOI) [139]: This metric measures the intelligibility of speech signals. It assesses the similarity between the temporal envelopes of the original and processed speech by correlating short-time segments from both. The metric yields a value between 0 and 1, with a higher value representing enhanced intelligibility.

- Log F0 Root Mean Square Error (F0 RMSE): This metric is a frame-level objective measure for evaluating the accuracy of predicted pitch contours. It calculates the RMSE between the logarithmic-scale F0 values of converted speech and those of the reference. It is defined as:

$$\text{F0 RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log f_0^{(\text{gen})}(i) - \log f_0^{(\text{ref})}(i)\right)^2}, \tag{2.16}$$

where N is the total number of voiced frames, and $f_0^{(\text{gen})}(i)$ and $f_0^{(\text{ref})}(i)$ are the F0 values of the generated and reference speech at frame i, respectively. A lower

F0 RMSE value quantifies a smaller F0 deviation, representing a more accurate pitch realization in the converted speech.

- Log F0 CORRelation (F0 CORR): This metric evaluates the similarity of the overall log F0 contours between converted speech and ground truth. It is computed as the Pearson correlation coefficient between the two log F0 sequences, where a higher value (closer to 1) indicates a stronger similarity in pitch contour shape and rhythm, while lower values reflect larger deviations.

2.4.2 Subjective Evaluation Metrics

This thesis employs two protocols of subjective listening tests to assess the perceptual performance of the proposed EL2SP systems: Mean Opinion Score (MOS) [140] for naturalness and speaker SIMilarity (SIM) for perceived speaker identity. All listening tests are conducted with native speakers of the language of the evaluated speech. The detailed listener setup and materials for each method are described in later chapters.

- MOS: Participants are requested to rate each presented speech sample in terms of how natural it sounds as normal human speech with good perceptual quality (i.e., perceived as natural rather than robotic or artificial). To provide clearer reference anchors, the original EL recordings and target normal recordings are also incorporated in the test, serving as the lower and upper bounds, respectively. The scoring criteria is given on the scale of 1 (completely unnatural) to 5 (completely natural). Specifically, on the evaluation page, each speech sample is accompanied by the following options: (1) *Bad*, (2) *Poor*, (3) *Fair*, (4) *Good*, and (5) *Excellent*, wherein higher scores indicate speech closer to the naturalness of normal speech.

- SIM: During the SIM test, listeners are presented with two speech samples at the same time, one from normal target speech and one from test speech. Similar to the setting of MOS, besides the samples for system, the original EL and

target normal speech are also included as test speech. Listeners are then asked to judge whether the two samples are spoken by the same speaker or not. Four-level response options are provided: *Definitely the same*, *Maybe the same*, *Maybe different*, and *Definitely different*. Results are summarized as the combined percentage of judgments falling into the former two categories (*Definitely the same* or *Maybe the same*), which reflects the degree of identity consistency. Thus, a higher SIM result indicates a higher speaker similarity to the target normal speaker.

2.5 Summary

This chapter provided the background knowledge and works related to most of this thesis. These discussions are essential for understanding the concepts and methods proposed in the subsequent chapters. Specifically, Section 2.1 presented an overview of fundamentals of seq2seq VC, covering its overall framework, major components, and network architecture, along with a review of related techniques in data augmentation and transfer learning. Section 2.2 introduced three categories of interference-robust VC methods, against which the proposed approach is compared to highlight its efficiency and effectiveness. Section 2.3 reviewed research on enhancing representation learning through auxiliary information, introducing the notion of multimodal fusion works. These works motivate the techniques developed in this thesis for improving mapping performance in EL2SP. Finally, in Section 2.4, the objective and subjective evaluation metrics employed in this thesis were summarized, which provide both quantitative and perceptual perspectives for assessing system performance and form the common basis for the experimental evaluations in Chapters 3–5.

The materials reviewed in this chapter lay the groundwork that is intrinsically linked to the training strategies and system designs presented in the remainder of this thesis. These foundational insights will be properly revisited in later chapters, pro-

viding the necessary background and practical references that support the proposed methods. In this way, readers are expected to approach the subsequent chapters with a clearer understanding of the advancements and contributions underlying the proposed methods of this thesis.

3 Pretraining and Fine-tuning Techniques for Electrolaryngeal Speech Enhancement

This chapter presents a series of training techniques based on sequence-to-sequence (seq2seq) Voice Conversion (VC) for ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP) under data-scarce conditions. While seq2seq VC is promising for EL2SP, it suffers performance degradation in the absence of sufficiently high-quality, parallel training data with the conventional pretraining–fine-tuning paradigm. Therefore, a multistage transfer learning framework is proposed here that extends beyond the conventional paradigm by incorporating complementary strategies at both the pretraining and fine-tuning phases. In particular, the pretraining phase contains a transfer learning step from Text-To-Speech (TTS) pretraining to VC, along with an innovative encoder adaptation training. The fine-tuning phase mainly comprises a two-stage scheme. These techniques effectively bridge the domain gap of upstream and downstream datasets and thus enhance the domain-specific generalization. Notably, aside from utilizing a publicly accessible TTS corpus and a minimal original EL2SP dataset, the proposed techniques incorporate low-quality Synthetic Data (SD), which is much easier to generate. This approach significantly alleviates the constraints caused by the

data-hungry property of seq2seq model and amplifies data efficiency. Experimental results presented in this chapter demonstrate the efficacy of the proposed techniques, with results showing dramatic improvements in both speech quality and intelligibility over conventional pretraining–fine-tuning baselines. This chapter also provides the detailed analyses across comparable systems with different techniques applied, further validating the contributions of the specific training components.

This chapter is organized as follows. Section 3.1 introduces the background and provides an overview of this work. Section 3.2 presents the proposed pretraining and fine-tuning methods. Section 3.3 describes the experimental evaluation settings, followed by the experimental results in Section 3.4. Section 3.5 concludes this chapter.

3.1 Introduction

This section presents the motivation and contributions of the proposed techniques. Despite the potential in EL2SP, applying seq2seq VC under low-resource conditions gives rise to two major technical challenges. As discussed in Subsection 1.3.1, the development of pretrained seq2seq VC models is hindered by (*P-1*) the scarcity of large-scale, parallel VC corpora, and (*P-2*) the severe domain mismatch between normal and EL2SP datasets further limits transferability. These issues undermine the effectiveness of conventional transfer learning. As will also be shown in later sections, directly fine-tuning a pretrained seq2seq model on the limited EL2SP data —while outperforming traditional ones [40]— still yields suboptimal performance. With all these said, this chapter aims to develop more tailored training strategies that better leverage the advantages of seq2seq modeling for EL2SP, while addressing the aforementioned two problems (*P-1* and *P-2*) of transfer learning under the same limited data conditions.

To address *P-1*, inspired by the Voice Transformer Network (VTN) [82], the previous work by the author of this thesis [45] followed a supervised method to construct a pretrained seq2seq VC model by leveraging a more easily accessible, high-quality TTS database. The method not only reduced the cost of training data but also allowed

the use of the parameters of the pretrained TTS model to effectively initialize the VC model.

After building the pretrained VC, data augmentation was applied as a potential strategy during fine-tuning to mitigate domain shifts mentioned in *P-2*. As stated in Subsections 1.3.4 and 2.1.5, data augmentation refers to the use of speech synthesizers to produce pseudo large-scale data for participating in direct VC training. This SD must be of high quality to ensure its feature distributions match closely with those of the small-scale original dataset. If SD cannot accurately simulate the feature distributions of the original dataset, the training will not generalize well. For instance, Biadsy *et al.* [72] developed a VC system to enhance speech produced by hearing-impaired speakers, using high-fidelity and -intelligibility synthetic voices produced by a high-performance TTS model. However, to ensure the SD quality, speech synthesizers such as TTS, with the network size like seq2seq VC, also require sufficient training data, which is still difficult to obtain. This problem is aptly termed the "chicken-and-egg problem" by Violeta *et al.* [44]. To avoid this problem, here, low-quality parallel SD is adopted to conduct EL2SP fine-tuning, given that constructing high-performance TTS models for synthesizing EL and normal speech with the minimal-resource dataset of EL2SP is infeasible. The argument here is that, despite significant distortions in low-quality SD, it preserves inherent structures from natural inputs, enabling seq2seq models to learn meaningful speech representations. On this basis, fine-tuning techniques are proposed here to assist the model in learning vast knowledge on an extended dataset by incorporating low-quality SD, while remedying the potential accuracy drifts caused by its imperfections.

However, current techniques can be regarded as an insufficient domain adaptation approach, aiming to alleviate domain-shifts between upstream normal speech and downstream EL2SP datasets through data augmentation, rather than fundamentally addressing their mismatches. Consequently, the pretrained model on normal speech pairs still encounter a significant adaptation gap when directly applied to EL2SP, mak-

ing it challenging to learn efficient cues from EL speech, even with fine-tuning. This prompts us to rethink better pretraining and fine-tuning methodologies to achieve a smoother representation learning process for seq2seq EL2SP.

Based on this, this chapter further enhances transfer learning between upstream and downstream datasets by refining both pretraining and fine-tuning. Building on the prior pretraining and fine-tuning pipeline, another level of encoder adaptation training is incorporated intermediately, which falls under the pretraining category in this study. As the prior pretraining only focuses on normal speech, this adaptation further refines the pretrained model, so that the domain representation is better adapted to the subdomain EL speech set used in the EL2SP fine-tuning. Note that low-quality data augmentation is introduced here to carry out this task. Because of the benefits obtained from low-quality synthetic speech for fine-tuning, we can anticipate that it will also be somewhat helpful in pretraining. Hereafter, the new EL2SP fine-tuning is conducted by integrating the updated module from encoder adaptation training while utilizing low-quality SD. The effectiveness of the proposed pretraining and fine-tuning methods are systematically investigated.

The contributions to this chapter are as follows:

- A unified training framework for seq2seq VC is proposed that addresses the issue of low-resource data while closing the domain-shift gap between pretraining and fine-tuning datasets through an encoder adaptation training that incorporates low-quality SD.
- Three systems are designed by combining two-stage fine-tuning strategies using low-quality SD with different levels of pretraining, each surpassing a typical *pretraining–fine-tuning* baseline for EL2SP [40].
- Experiments are conducted with different setups for two EL2SP datasets: (1) Minimal-resource parallel dataset, and (2) More severe semi-parallel dataset. All the experimental results show that the systems with the simultaneous introduction of new encoder adaptation training with fine-tuning strategies achieve best re-

sults in terms of naturalness and intelligibility, demonstrating that the encoder adaptation training can improve the pretrained model for the EL data domain.

- Unlike most systems with demanding requirements for training data, the proposed methodology focuses on the more practical establishment scenario for EL2SP based on low-quality SD. Grouped experiments with different amounts of low-quality SD reveal that the model performance is positively correlated with the amount of low-quality SD, further verifying its feasibility.

3.2 Pretraining and Fine-tuning Methods

As discussed in Sections 2.1, and 3.1, existing methods face two problems when addressing the data-hungry property of seq2seq VC: (1) Neglecting the negative impact on transfer learning when the upstream task is quite different from the downstream task, and (2) Depending on the quality of external modules or data augmentation. At the core of this chapter's study is the idea of proposing novel pretraining and fine-tuning strategies to improve the performance of the seq2seq EL2SP system, requiring only a small initial dataset, even in the presence of non-parallel portions. To facilitate understanding of the proposed techniques, Figure 3.1 provides a simplified conceptual illustration before introducing the details. Aside from the similar techniques of pretraining on normal corpora and fine-tuning on EL2SP data to the prior work by the author of this thesis [45], an encoder adaptation training tailored to feature distributions of EL speech is introduced. It is designed to improve the generalization and minimize the learning gap in comprehending EL speech owing to domain shifts, thus for more refined EL2SP fine-tuning.

Specifically, the pretraining techniques are divided into two parts. First, the pretrained TTS is utilized to build the pretrained seq2seq model (Subsection 3.2.1). Subsequently, the encoder adaptation training for EL speech based on low-quality EL SD is carried out (Subsection 3.2.3). On the other hand, the effective two-stage fine-tuning strategy using low-quality SD from the prior work [45] is extended to improve and

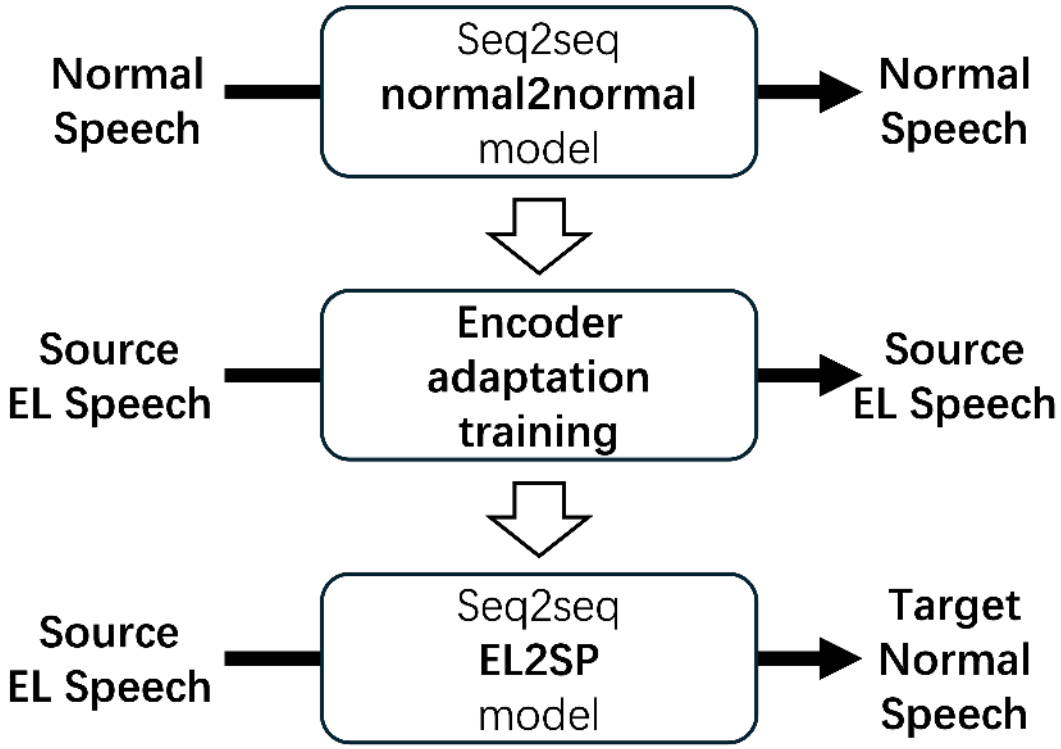


Figure 3.1: Concept of smoother transfer learning from sequence-to-sequence (seq2seq) Voice Conversion (VC) pretraining to seq2seq ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP) fine-tuning with encoder adaptation training.

stabilize the model with a minimal EL2SP dataset (Subsection 3.2.4). Notably, the TTS models are fine-tuned by an extra fine-tuning stage, to supply the essential components and low-quality SD for the implementation of the unified pretraining and fine-tuning techniques (Subsection 3.2.2). Finally, various steps are integrated to develop different systems (Subsection 3.2.5). The architecture containing all pretraining and fine-tuning stages is depicted in Figure 3.2. More details are elaborated on the following subsections.

3.2.1 Pretraining of a Normal Sequence-to-sequence (Seq2seq) Model

The Transformer-based TTS pretraining is adopted for constructing a one-to-one VTN. Seq2seq VC encodes the source speech into hidden representations, which are

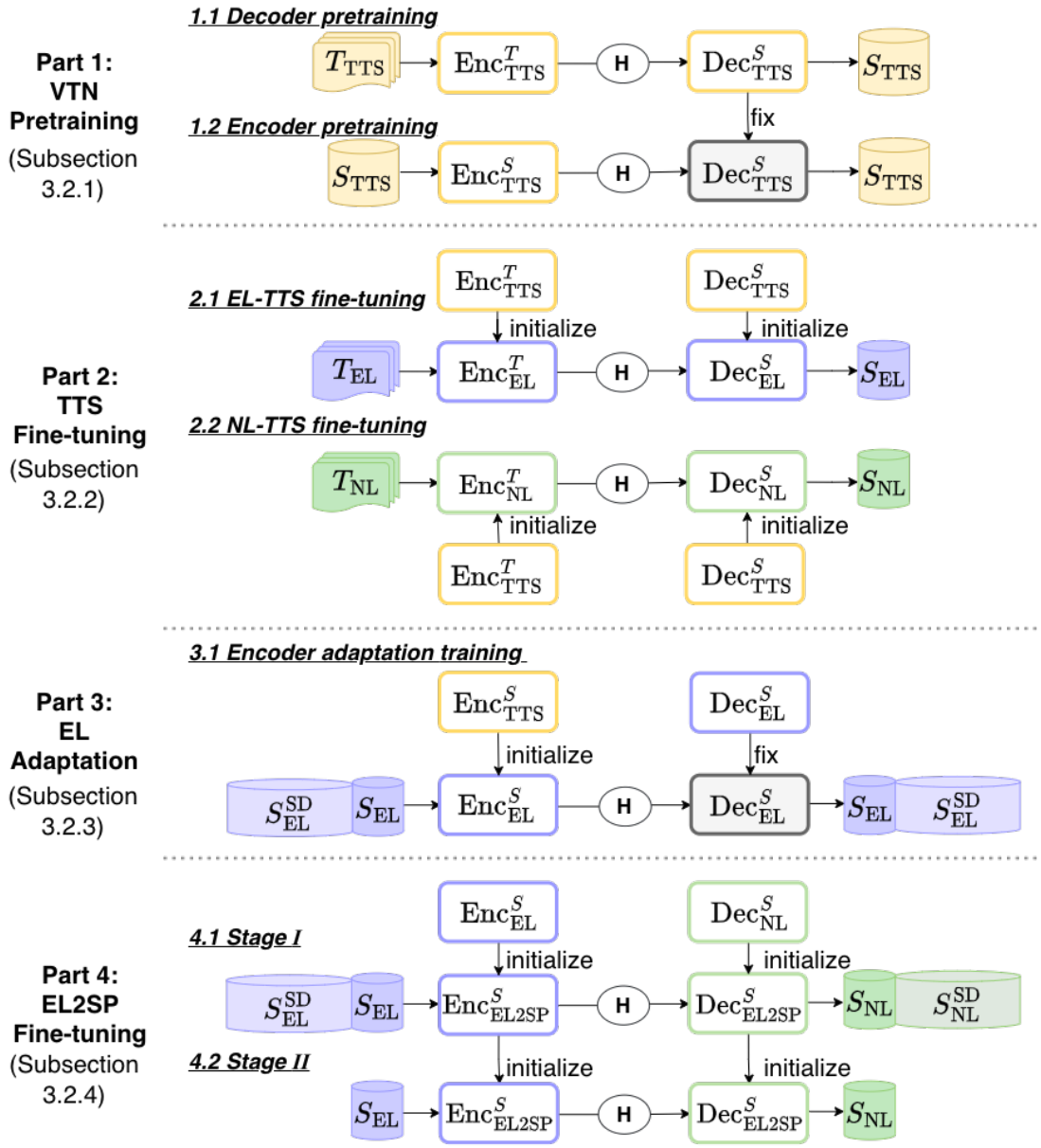


Figure 3.2: Overall system consisting of four parts constructed by multiple pretraining and fine-tuning techniques. Part 1: Pretraining of Voice Transformer Network (VTN) using a normal speech corpus. Part 2: ElectroLaryngeal (EL) and normal Text-To-Speech (TTS) fine-tuning. Part 3: Encoder adaptation training for EL speech incorporating low-quality EL SD. Part 4: EL2SP fine-tuning incorporating low-quality parallel Synthetic Data (SD). Parts 1 to 4 correspond to the descriptions detailed in Subsections 3.2.1 to 3.2.4, respectively.

then decoded into the target speech. Essentially, the encoding of VC serves to eliminate speaker identity from the source speech, providing the decoder with speaker-independent hidden representations. In contrast, TTS involves a text-to-speech mapping process, which more readily generates pure linguistic hidden representations

from text for speech production. In summary, the hidden representations derived from TTS and the mechanism for decoding them into speech are beneficial for VC in generating high-fidelity speech. Moreover, the TTS pretraining only requires an arbitrary single-speaker corpus and the corresponding text. This naturally induces the transfer of capabilities from TTS to VC during pretraining stages.

Let $D_{\text{TTS}} = \{T_{\text{TTS}}, S_{\text{TTS}}\}$ denote a large normal TTS corpus, where T_{TTS} and S_{TTS} represent the text and speech sets, respectively. Part 1 in Figure 3.2 illustrates a specific VTN pretraining architecture comprising two substages for bringing the base-prior parameters to VC. Details are as follows:

- *Decoder pretraining*: A typical TTS training from scratch is first conducted using D_{TTS}. The updated text encoder $\text{Enc}^{T}_{\text{TTS}}$ can effectively encode the text set into hidden representations of pure linguistic information, enabling the speech decoder $\text{Dec}^{S}_{\text{TTS}}$ to exhibit a robust capability to map between various speech features and linguistic information.

- *Encoder pretraining*: An autoencoder is then involved as the speech encoder $\text{Enc}^{S}_{\text{TTS}}$ to substitute the original $\text{Enc}^{T}_{\text{TTS}}$ and still trained with D_{TTS}, whose speech set S_{TTS} is used as both input and target datasets. Here, the parameters of pretrained $\text{Dec}^{S}_{\text{TTS}}$ are frozen and only $\text{Enc}^{S}_{\text{TTS}}$ is updated by minimizing the reconstruction loss. In this manner, the hidden representations and the capability of the pretrained decoder are not only inherited, but can also be leveraged to compel $\text{Enc}^{S}_{\text{TTS}}$ to accurately learn to extract analogous linguistic hidden representations from speech rather than text.

Note that *Decoder pretraining* can extract rich linguistic hidden representations from a well-trained text encoder. During *Encoder pretraining*, fixing the decoder can establish a pre-defined target with the inherited hidden representations, which calibrates $\text{Enc}^{S}_{\text{TTS}}$ to learn acoustic features similarly to $\text{Enc}^{T}_{\text{TTS}}$ encoding text. Also note that the pretraining outlined above utilizes the same normal corpus because it is more

easily obtained, and the hidden representations of normal VC can provide necessary linguistic knowledge for EL2SP.

3.2.2 Text-To-Speech (TTS) Fine-tuning and Parallel Synthetic Data (SD) Generation

The procedure illustrated in Part 2 in Figure 3.2 aims to provide data augmentation and the essential components required for the subsequent parts. Consequently, both TTS models for normal and EL speech are trained. However, the initial source EL corpus $D_{\text{EL}} = \{T_{\text{EL}}, S_{\text{EL}}\}$ and the target normal corpus $D_{\text{NL}} = \{T_{\text{NL}}, S_{\text{NL}}\}$ are too small to be trained from scratch. Given such circumstances, the pretrained TTS model is fine-tuned, which is obtained through *Decoder pretraining* described in Subsection 3.2.1, on the basis of the two corpora, thus yielding the corresponding TTS models, TTS_{EL} and TTS_{NL}, respectively.

Finally, an external text set is fed into the two systems simultaneously to generate EL and normal parallel synthetic speech, denoted by $D_{\text{PSD}} = \{S^{\text{SD}}_{\text{EL}}, S^{\text{SD}}_{\text{NL}}\}$. Note that $S^{\text{SD}}_{\text{EL}}$ and $S^{\text{SD}}_{\text{NL}}$ are of low-quality due to the poor performance of TTS_{EL} and TTS_{NL}, which is constrained by the minimal original dataset. In addition, the speech decoders of TTS_{EL} and TTS_{NL}, denoted by $\text{Dec}^{S}_{\text{EL}}$ and $\text{Dec}^{S}_{\text{NL}}$, can be utilized in Parts 3 and 4, respectively.

3.2.3 Encoder Adaptation Training

As shown in Part 3 in Figure 3.2, the objective here is to conduct the adaptation training to enable the encoder to recognize EL speech. Similar to *Encoder pretraining*, the input EL speech is reconstructed in an autoencoder style. Since EL speech is very limited, the low-quality EL SD $S^{\text{SD}}_{\text{EL}}$ generated by the fine-tuned EL TTS model in Part 2 is first combined with the original S_{EL} for both input and output datasets. The intuition is that the large low-quality $S^{\text{SD}}_{\text{EL}}$ possesses essential features close to those of natural EL speech, which can enhance the adaption training. Meanwhile, $\text{Enc}^{S}_{\text{TTS}}$ of Part 1

is considered to obtain the pretrained weights for recognizing normal human speech via large-scale D_{TTS}, and $\mathrm{Dec}^{S}_{\mathrm{EL}}$ in Part 2 can preliminarily synthesize EL speech via the minimal D_{EL}. On the basis of the above, the encoder and decoder are initialized using the parameters of $\mathrm{Enc}^{S}_{\mathrm{TTS}}$ and $\mathrm{Dec}^{S}_{\mathrm{EL}}$, respectively, instead of being trained from scratch. Next, $\mathrm{Dec}^{S}_{\mathrm{EL}}$ is fixed so that the adaptation training concentrates on updating the encoder to obtain $\mathrm{Enc}^{S}_{\mathrm{EL}}$ for EL speech.

3.2.4 Two-stage ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP) Fine-tuning

As present in Part 4 in Figure 3.2, the fine-tuning technique is proposed to train the desired EL2SP model that contains the encoder $\mathrm{Enc}^{S}_{\mathrm{EL2SP}}$ and the decoder $\mathrm{Dec}^{S}_{\mathrm{EL2SP}}$ using different initialization parameters. On the one hand, $\mathrm{Dec}^{S}_{\mathrm{NL}}$ in Part 2 is used to initialize $\mathrm{Dec}^{S}_{\mathrm{EL2SP}}$. Despite the trivial benefits on performance compared with initializing with $\mathrm{Dec}^{S}_{\mathrm{TTS}}$ owing to the minimal fine-tuning dataset D_{NL}, it can somewhat expedite the model convergence because S_{NL} is the original target set of EL2SP. On the other hand, the parameters of $\mathrm{Enc}^{S}_{\mathrm{EL}}$ from encoder adaptation training are used as valid a *priori* information to initialize $\mathrm{Enc}^{S}_{\mathrm{EL2SP}}$, thus achieving higher transferability and performance for EL2SP.

Let $D_{\mathrm{EL2SP}} = \{S_{\mathrm{EL}}, S_{\mathrm{NL}}\}$ represent original natural speech pairs. During Stage I, D_{EL2SP} and D_{PSD} are combined as the training set to obtain the EL2SP model, which is denoted as $\mathrm{VTN}_{\mathrm{1st}}$. Although D_{PSD} can model the EL and normal speech in terms of intonation and speaker identity, it contains some misinformation owing to the limited performance of the TTS models, which may inhibit EL2SP from building perfect hidden representations. The expection here is that the limited natural dataset can further correct the training. Thus, Stage II is set up to obtain the final EL2SP model, $\mathrm{VTN}_{\mathrm{2nd}}$, during which the parameters of $\mathrm{VTN}_{\mathrm{1st}}$ are initialized and D_{EL2SP} is re-input for training.

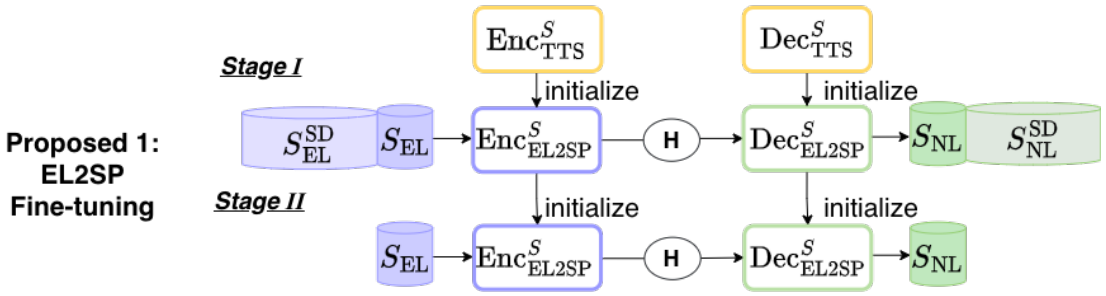


Figure 3.3: EL2SP fine-tuning of the Proposed 1 system using pretrained VTN.

3.2.5 Proposed EL2SP Systems

On the basis of the pretraining and fine-tuning techniques in the above subsections, three systems named Proposed 1, Proposed 2, and Proposed 3, are proposed here.

- *Proposed 1*: In this system, the pretrained VTN of Part 1 is used as the initialization for EL2SP fine-tuning. The TTS models are utilized for generating parallel SD, which is pooled together with the original dataset for the two-stage fine-tuning, as shown in Figure 3.3. During this process, the knowledge transfer from *VTN pretraining* to EL2SP relies solely on the fine-tuning techniques in the absence of encoder adaptation training.

- *Proposed 2*: This system is developed by exploiting all the pretraining and fine-tuning settings in Figure 3.2, where the encoder adaptation training is added to pretraining stages.

- *Proposed 3*: Similarly to *Proposed 2*, this system follows all the pretraining and fine-tuning techniques. The key difference from *Proposed 2* lies in *TTS fine-tuning*, as shown in Figure 3.4, where the text encoders remain constant, whereas only the speech decoders for EL and normal speech are updated to preserve the pretrained hidden representations for initializing *EL2SP fine-tuning*. Note that *VTN pretraining* with a larger natural dataset can already build compact hidden representations with rich linguistic patterns, which are essential for high-fidelity and highly intelligible converted speech during *EL2SP fine-tuning*. As the overall

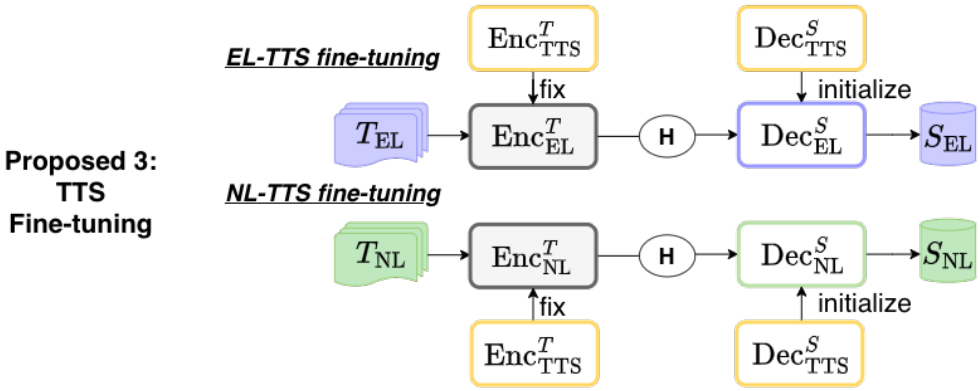


Figure 3.4: EL and normal TTS fine-tuning of the Proposed 3 system by fixing text encoders.

pretraining is a multistage process, and the low-quality SD is added for training, it is possible that the original information obtained from *VTN pre-training* is diminished and hurts accuracy on the fine-grained downstream dataset. So, the specific modules are selectively frozen and only the entire model is updated during *EL2SP fine-tuning* to better integrate the pretrained knowledge with the EL2SP fine-tuning dataset.

3.3 Experimental Evaluation Settings

This section presents the experimental evaluation settings of this chapter, including the datasets, implementation details, waveform synthesis modules, baseline systems, and evaluation metrics.

3.3.1 Datasets and Implementation

3.3.1.1 Dataset for VTN Pretraining and SD Generation

The pretrained TTS model and pretrained VTN were sequentially built according to *Decoder pretraining* and *Encoder pretraining* as described in Subsection 3.2.1, using a normal Japanese corpus named JSUT database [141] recorded by a female speaker. The training set contained 7,296 utterances, roughly 10 hours long. In addition, the text data of the training set was chosen for generating external parallel SD.

3.3.1.2 Original EL2SP Datasets

Two original small-scale Japanese datasets were constructed to evaluate the proposed methods. One contained parallel pairs of EL and normal corpora recorded from a healthy male speaker using an electrolarynx and his normal voice, denoted by SimuEL and NormSP, respectively. Each pair contained 413 utterances, totaling 20 minutes. The EL corpus from the other dataset was contributed by an actual male laryngectomee, denoted by RealEL. Since it was unable to record the normal corpus from this laryngectomee, NormSP was shared as the reference target speech. Note that RealEL contained only 200 sentences within 10 minutes, whose utterance contents were only part of NormSP, so this dataset belonged to a semi-parallel corpus.

These two datasets determine two corresponding experimental scenarios for all the proposed methods. In Case 1, the objective was to convert SimuEL to NormSP (SimuEL–NormSP), simulating the scenario where the normal corpus of the patient was available. In Case 2, the objective was to convert RealEL to NormSP (RealEL–NormSP), which represented a more practical but severe case. In each case, 20 utterances from the respective dataset were used as the development set, another 40 as the evaluation set, and the remainder as the training set. To utilize all feasible data in RealEL–NormSP, the original corpus was first supplemented with the corresponding EL SD to address the non-parallel part of the original corpus. On the grounds that the datasets of two cases were different, different evaluation sets were built for simplicity.

3.3.1.3 Implementation Details

The proposed systems were entirely implemented using the open-source ESPnet toolkit [96], [142], where all speech data from pretraining to fine-tuning were resampled at 24 kHz, and the 80-dimensional mel filterbanks with 2,048 Fast Fourier Transform (FFT) points and a 300-point frame shift were used to extract the acoustic features. VTN for constructing the EL2SP and TTS models for generating SD shared

similar base configurations using six-layer blocks with four attention heads in the encoder and decoder. The input feature sequence of the TTS model contained phoneme and pause information. Meanwhile, the learning rate was set to 0.001, while the LAMB optimizer [143] for VC and the Noam optimizer [87] for TTS were employed, respectively.

3.3.1.4 Waveform Synthesis Modules

For a rapid turnaround, the Parallel WaveGAN (PWG) neural vocoder [144] was used to efficiently reconstruct the waveform of SD and the EL2SP outputs. The corresponding speaker-dependent PWGs were trained separately using SimuEL, NormSP, and RealEL. Note that the PWG vocoder for reconstructing NormSP was trained from scratch, whereas SimuEL and RealEL were pretrained from an additional EL set with 1,000 utterances, which were simulated by another healthy speaker.

3.3.1.5 Baseline Systems

Two types of baselines were prepared, non-seq2seq and seq2seq, each consisting of two systems corresponding to Cases 1 and 2. The non-seq2seq baseline systems were implemented using a Convolutional Long-short term Deep Neural Network (CLDNN)-based model proposed by Kobayashi and Toda [2], which is a typical frame-wise EL2SP approach. These baseline systems were trained from scratch using only the parallel pairs from the original EL2SP corpus, i.e., 353 pairs for Case 1 and 140 pairs for Case 2, as they could not leverage pretraining on a TTS database. These baselines served to reflect the architectural advantage of the seq2seq modeling under low-resource settings.

The seq2seq baseline systems were set up using the typical Transformer-based architecture, as introduced in Subsection 2.1.3, following the methodology by Yen *et al.* [40]. To ensure a fair comparison, they shared the identical framework and VTN pretraining process to those of the proposed systems, but omitted encoder adaptation training for EL speech and low-quality SD introduction during EL2SP fine-tuning. In-

stead, they directly followed the conventional *pretraining–fine-tuning* paradigm, where fine-tuning used the same parallel data as in the non-seq2seq baselines and Stage II of the proposed systems.

The reason for adopting both baselines was twofold. First, comparing the non-seq2seq and seq2seq baselines under the same data conditions allows us to confirm that the seq2seq ones offered superior performance for EL2SP regarding speech quality, naturalness, and intelligibility, as preliminarily indicated by Yen *et al.* [40]. Second, focusing on benchmarking against the stronger seq2seq baselines allows us to comprehensively quantify the contributions of the proposed training techniques, particularly the impact of using low-quality SD toward improving the proposed EL2SP systems.

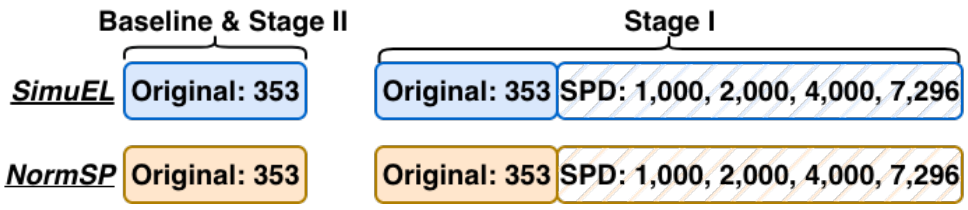
3.3.2 Evaluation Metrics

3.3.2.1 Objective Evaluation

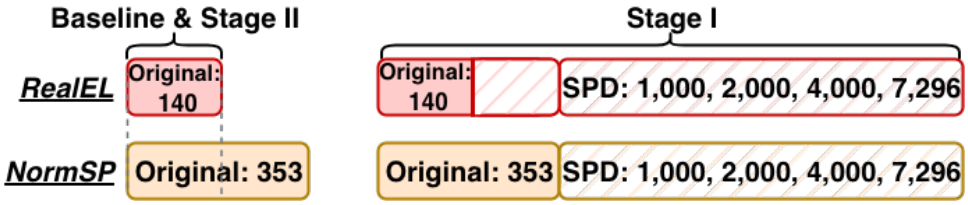
Two objective metrics, Mel-Cepstrum Distortion (MCD, in dB) and Character Error Rate (CER, in %), as described in Subsection 2.4.1, were used to assess the proposed EL2SP systems in terms of general spectral quality and the intelligibility. Moreover, since the research in this chapter is closely tied to SD, CER was used as a rating criterion to examine the quality of SD. Therefore, a lower CER value could be viewed as an indicater of higher intelligibility accuracy for the converted/synthetic speech. Here, the ASR engine utilized for calculating CER had a conformer-based architecture [145] pretrained using the Japanese LaboroTV database [18] and fine-tuned as by Violeta *et al.* [146] to adapt it to EL speech. It could recognize all original EL (combined with RealEL and SimuEL) and normal speech with the best CERs of 14.7% and 6.9%, respectively, which can be regarded as the upper bounds of ASR results to reflect the quality of SD.

3.3.2.2 Subjective Evaluation

For the subjective tests, Mean Opinion Score (MOS) as defined in Subsection 2.4.2, was performed to assess the naturalness of each presented speech sample. In Cases 1



(a) Case 1: SimuEL–NormSP.



(b) Case 2: RealEL–NormSP.

Figure 3.5: Training datasets for the baseline systems and Stages I and II of the proposed systems for Cases 1 and 2. 1,000, 2,000, 4,000 and 7,296 represent the screened datasizes of SD used for the experiments.

and 2, fifteen randomly selected utterances from each system were included in the naturalness tests. More than thirty Japanese participants were recruited. Audio samples are available online[1].

3.4 Experimental Results

The Proposed 1, 2, and 3 systems share a similarity in that they all focus on the development of pretraining and fine-tuning techniques with TTS-generated SD. Therefore, it is essential to study the impact of SD on these systems. In this context, a distinct system for each method was specifically developed using original datasets with different amounts of SD. Note that the term *datasize* refers to the volume of the data in the following discussion.

3.4.1 Quality of External SD

As described in Subsection 3.2.2, SD was generated using the fine-tuned TTS models. Note that the TTS models for generating SD are not universal between Cases 1 and 2 owing to the different training sets used. In each case, Proposed 1 and 2 adopted

[1]See https://silenticymoon.github.io/EL2SP-demo/ for samples.

Table 3.1: CER values in % characterizing the quality of EL and Normal SD with datasizes of 1,000, 2,000, 4,000, and 7,296 for building the Proposed 1, 2, and 3 systems. SD-(1,000–7,296) represents the SD size and the corresponding system in Case 1: SimuEL–NormSP.

Systems	**Proposed 1 and 2**		**Proposed 3**	
External SD	EL	Normal	EL	Normal
SD-1,000	35.3	16.4	37.7	19.2
SD-2,000	39.4	20.0	42.2	23.3
SD-4,000	44.5	25.6	46.7	28.6
SD-7,296	53.0	34.8	55.5	37.2

Table 3.2: CER values in % characterizing the quality of EL and Normal SD with datasizes of 1,000, 2,000, 4,000, and 7,296 for building the Proposed 1, 2, and 3 systems. SD-(1,000–7,296) represents the SD size and the corresponding system in Case 2: RealEL–NormSP.

Systems	**Proposed 1 and 2**		**Proposed 3**	
External SD	EL	Normal	EL	Normal
SD-1,000	49.5	17.2	53.8	19.9
SD-2,000	54.8	21.0	60.0	24.2
SD-4,000	61.5	25.7	67.1	29.2
SD-7,296	71.7	33.7	75.6	36.8

identical TTS models for generating EL SD and normal SD, whereas Proposed 3 trained exclusive TTS models using a unique TTS fine-tuning technique described in Subsection 3.2.5. For each method, four datasize-driven systems were constructed using 1,000, 2,000, 4,000, and 7,296 external SD. Figure 3.5 demonstrates the EL2SP training datasets of the baseline and proposed systems for both cases. Encoder adaptation training in Proposed 2 and 3 was carried out for each datasize using their respective EL SD identical to Stage I of EL2SP fine-tuning. An ASR-based filtering system was built. Specifically, according to CERs from all external SD (7,296), the system filtered out and selected the parallel training pairs with the highest quality at each datasize, so as to reduce the misinformation for training. Tables 3.1 and 3.2 show the calculated CERs of selected SD for each datasize in Cases 1 and 2, respectively, as the validation of the SD quality.

Several intriguing properties of SD quality were observed. It is evident that no matter which system was examined, the quality of EL SD and normal SD was poor because of the limited original dataset for TTS training. This was especially true for

Table 3.3: Comparison of CLDNN-based non-seq2seq baselines and seq2seq baselines for Case 1: SimuEL–NormSP, and Case 2: RealEL–NormSP.

Systems	Case 1		Case 2	
	MCD [dB]	CER [%]	MCD [dB]	CER [%]
CLDNN Baseline [2]	10.31	55.8	10.43	41.7
Seq2seq Baseline [40]	6.37	51.4	7.17	41.3

EL SD, which revealed the significant challenge of transfer learning from an upstream normal database to a downstream EL set.

When examining each proposed system individually, the quality of EL SD and normal SD progressively deteriorated with increasing datasize owing to the inclusion of lower-quality SD during data screening in Cases 1 and 2. For instance, the CER of EL SD for Proposed 1 and 2 sharply increased from 49.5 % to 71.7 %, as shown in Table 3.2. By comparing both cases under the same datasize, we can see that the quality of normal SD was similar, whereas the EL SD in Table 3.2 was markedly inferior to that in Table 3.1, since the original EL set for RealEL (140) was over half size smaller than that for SimuEL (353), resulting in a worse performance of the EL TTS models. The quality discrepancy for EL SD in two cases gradually enlarged as the datasize increased. At the datasize of 7,296, the quality difference of EL SD in CER for Proposed 1 and 2 reached 18.7 percentage points (53.0 % in Table 3.1 versus 71.7 % in Table 3.2) and for Proposed 3 reached 20.1 percentage points (55.5 % in Table 3.1 versus 75.6 % in Table 3.2).

We can see that the SD quality in Proposed 3 was consistently lower than that in Proposed 1 and 2 across all datasizes, owing to the fact that the TTS fine-tuning for Proposed 3 only updates the speech decoder to maintain hidden representations, thereby limiting the performance of entire TTS models.

3.4.2 Objective Evaluation Results

To establish a reference point for comparison, first, the two baseline systems were compared before presenting the results of the proposed systems. As shown in Table 3.3,

the seq2seq baselines consistently outperformed the non-seq2seq counterparts across Cases 1 and 2, in terms of both MCD and CER. Therefore, the seq2seq baselines are adopted as the primary reference systems in the subsequent experiments.

Next, by combining with the observations for SD quality, the performance characteristics of the proposed systems are further studied here. The results are shown in Tables 3.4 and 3.5 for Cases 1 and 2, respectively. Note: "S2S base." in these two tables denotes the seq2seq baseline system.

Comparison with seq2seq baseline: The proposed methods are initially compared with the seq2seq baseline systems. As mentioned above, the SD used in Case 1 is of low quality. Nevertheless, Table 3.4 qualitatively demonstrates that the overall results from Stages I and II for Proposed 1, 2, and 3 were all significantly better than the seq2seq baseline in terms of speech quality and intelligibility. In Case 2, despite using a much poorer EL SD, the three proposed systems still surprisingly outperformed the corresponding baseline with notable margins for MCD and CER, as shown in Table 3.5. In particular, the results of the seq2seq baselines for both cases were well below those of the counterparts in Stage I among the three proposed systems using the lowest-quality 7,296 SD. Taken together, it can be concluded that the proposed pretraining and fine-tuning techniques using parallel SD and the original dataset can assist the system to build rich-linguistic hidden representations from Stage I.

Impacts with different SD sizes: Looking at the results of individual methods across different datasizes reveals an interesting finding for Case 1. When the SD size increased from 1,000 to 7,296, albeit with a concurrent decline in SD quality, the outcomes of Stage I consistently improved in terms of both MCD and CER. This trend is observed in 23 out of all 24 results among the three proposed methods. As expected, Stage II yielded better results by conducting the same training procedure to models derived from Stage I with different SD datasizes. These results underscore the effectiveness of imperfect SD in larger datasizes for model performance, and Stage II can minimize the error from SD.

Table 3.4: Objective evaluations for the Proposed 1, 2, and 3 systems with different SD datasizes and the seq2seq baseline for Case 1: SimuEL–NormSP.

Task		**Proposed Systems**													
		Proposed 1				**Proposed 2**				**Proposed 3**				**S2S base.** [40]	
Original	Systems	Stage I		Stage II		Stage I		Stage II		Stage I		Stage II		—	
		MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]
SimuEL – NormSP	SD-1,000	5.95	42.1	5.93	39.4	5.93	36.6	5.82	36.7	5.88	40.5	5.82	38.9	6.37	51.5
	SD-2,000	5.90	40.7	5.82	37.8	5.88	36.7	5.80	36.2	5.87	38.6	5.77	37.6		
	SD-4,000	5.87	37.4	5.78	36.9	5.83	35.5	5.72	34.0	5.77	38.4	5.73	36.8		
	SD-7,296	5.78	36.3	5.77	34.7	5.77	35.4	5.70	33.6	5.69	35.9	5.61	34.4		

Table 3.5: Objective evaluations for the Proposed 1, 2, and 3 systems with different SD datasizes and the seq2seq baseline for Case 2: RealEL–NormSP.

Task		**Proposed Systems**													
		Proposed 1				**Proposed 2**				**Proposed 3**				**S2S base.** [40]	
Original	Systems	Stage I		Stage II		Stage I		Stage II		Stage I		Stage II		—	
		MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]	MCD [dB]	CER [%]
RealEL – NormSP	SD-1,000	6.66	26.9	6.59	24.5	6.45	26.0	6.36	21.2	6.57	25.8	6.14	22.5	7.17	41.3
	SD-2,000	6.48	27.7	6.41	25.3	6.38	26.5	6.33	20.8	6.22	26.5	6.08	25.1		
	SD-4,000	6.28	26.3	6.18	21.9	6.26	25.4	6.20	17.7	6.24	25.7	6.05	23.6		
	SD-7,296	6.26	29.1	6.24	23.3	6.27	25.3	6.22	22.4	6.30	25.3	5.88	21.1		

Stage I in Case 2 does not fully replicate the finding in Case 1. For Proposed 1, MCD continuously decreased as SD datasize increased from 1,000 to 7,296. Furthermore, the optimization of CER was not apparent and remained at a steady level. When using all SD (7,296), MCD turned out to be the best, while CER increased compared to the scenario of 4,000. For Proposed 2 and 3, the improvement tended to fluctuate slightly with increasing datasize, but the most minimal MCD and CER were always observed when pooling the datasize of 4,000 or all SD into the training of Stage I. Such low-intelligibility SD in Case 2 would affect the proposed systems in converting accurate utterance contents. However, a sufficiently large-scale SD can still positively improve both speech quality and intelligibility, especially for Proposed 2 and 3. Next, a consistently higher performance was seen during Stage II for each datasize, with larger datasizes showing nontrivial improvements, e.g., for SD at 4,000, the CER decreased from 26.3 % to 21.9 % for Proposed 1 and from 25.4 % to 17.7 % for Proposed 2.

The analysis above confirms the robustness of the proposed systems. In both cases, despite the unsmooth trend of several datasizes in Case 2, extending the scale of low-quality SD can provide richer knowledge to improve the performance of Stage I. Stage II maximizes the advantages of large-scale SD, consequently having a negligible negative impact on CER.

Comparison with Proposed 1: Let us analyze the performance characteristics of Proposed 1 against those of Proposed 2 and 3, as the latter two include encoder adaptation training for EL speech. Intuitively, Proposed 2 in both cases outperformed Proposed 1 during Stage I for each datasize. In addition, compared with the advantage of MCD, Proposed 2 showed a more pronounced superiority in terms of CER. This suggests that the encoder adaptation training using EL SD facilitates the understanding of the model for the linguistic information from EL speech. Unsurprisingly, Proposed 2 gained obviously better results in Stage II than Proposed 1 for individual datasizes. Examining the best outcomes of methods both in Case 2, Proposed 2 showed a similar

MCD with Proposed 1, but achieved a much lower CER of 17.7, which stands out as the best among all three proposed methods.

Similar conclusions could be observed for Stages I and II when comparing Proposed 1 and 3. In both cases, Proposed 3 performed better than Proposed 1 more often than not and won in 14 of the 16 pairs compared, whereas the differences in the remaining two pairs were insignificant. Note that Proposed 3 uses the lower-quality SD listed in Tables 3.1 and 3.2 compared with the other systems for encoder adaptation training and EL2SP fine-tuning. At a sufficient scale, despite much misinformation in low-quality EL SD, we can say that it still contains the typical features of EL speech and speaker information, which can be leveraged by the pretrained model to initially enable the recognition of EL speech during encoder adaptation training.

In summary, the higher performance characteristics of Proposed 2 and 3 demonstrated a higher data efficiency than Proposed 1. Even using SD of lower quality, Proposed 3 still exhibited enhanced performance, validating the high robustness of the encoder adaptation training and fine-tuning techniques.

Comparison between Proposed 2 and 3: From the overall results, Proposed 2 and 3 showed some overlapping properties, which particularly can benefit considerably from increasing the SD size. Furthermore, Stage II aids both methods in achieving the optimal results. On the other hand, besides showing advancements and sharing the same optimization direction for both MCD and CER compared to Proposed 1, Proposed 2 and 3 also revealed distinctive advantages. In both cases, the CERs of Proposed 2 were mainly superior to those of Proposed 3, whereas Proposed 3 demonstrated a greater potential in terms of MCD, and most of the major improvements were particularly concentrated on the several final outcomes of Case 2. Given that MCD and CER determine different aspects, the disparity in results for the two metrics can be viewed as a specific trade-off by Proposed 2 and 3 between (1) whether EL speech is accurately converted into a natural pronunciation that closely resembles the target normal voice, and (2) whether there are inconsistencies in the pronounced

contents of converted speech.

In view of this, the methodology of Proposed 3 is analyzed here, which seeks to preserve the hidden representations of natural speech from a normal TTS database during pretraining. These fixed representations reduce the influence of downstream EL speech features/pronunciation rules, facilitating the model to establish a mapping that leans towards the pronunciation of normal target voice during EL2SP fine-tuning. On the other hand, keeping hidden representations necessitates that TTS fine-tuning focuses solely on updating the normal speech decoder, potentially providing EL2SP fine-tuning with parameters that are more beneficial for reconstructing the acoustic properties of the target normal speech. Thanks to these, Proposed 3 achieved lower MCDs, reflecting better-sounding converted voices that more closely match the target. However, to save the hidden representations, Proposed 3 must use SD with higher CER at the same time, where more errors in SD continuously accumulate during the downstream training processes, negatively impacting the recognition performance of the EL encoder as well as the effectiveness of the EL2SP fine-tuning. These compromises, stemming from SD quality, inevitably restrict further optimization of intelligibility, which manifest in higher CERs. In contrast, Proposed 2 adopts a more direct training method that can generate relatively better SD for assisting in the conversion to a higher-intelligibility speech. Although the hidden representations are changed during pretraining, potentially remaining some EL pronunciation properties, they generally do not impair the intelligibility, while the converted speech may sound less natural, as indicated by higher MCDs.

To further verify the analysis, a new system named *Proposed Hybrid* is designed, which employs the two-stage fine-tuning for the pretrained model of Proposed 3 while using the higher-quality SD from Proposed 2. Experiments were conducted under Case 2, and the results are documented in Table 3.6. Inspection suggests the performance of Proposed Hybrid was generally well above the overall results of Proposed 2 and 3 in Table 3.5. The advantage over Proposed 2 is a clear indicator of

Table 3.6: Objective evaluations for the Proposed Hybrid system with different SD datasizes for Case 2: RealEL–NormSP.

Task		**Proposed Hybrid**			
Original	Systems	Stage I		Stage II	
		MCD [dB]	CER [%]	MCD [dB]	CER [%]
RealEL – NormSP	SD-1,000	6.39	24.2	5.93	20.7
	SD-2,000	6.11	24.8	6.06	20.6
	SD-4,000	6.03	24.8	5.83	21.0
	SD-7,296	6.09	24.0	5.86	20.2

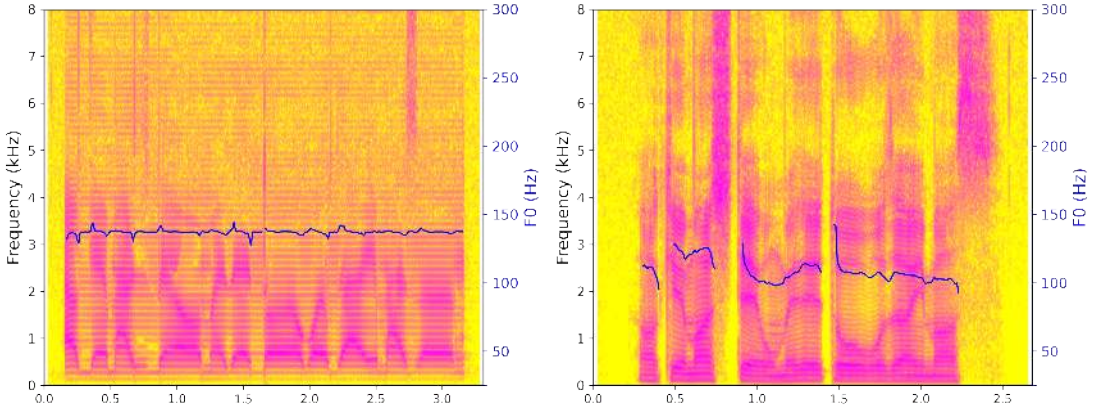
the effectiveness of the method that preserves the hidden representations. Moreover, the advancement of Proposed Hybrid in CER reflects the effect of higher-quality SD, which can compensate for the negative impact on the intelligibility of the converted speech.

3.4.3 Spectrogram Analysis

Figure 3.6 visualizes the performances of the proposed systems by plotting the spectrograms and F0 contours of a converted speech. We can see the following: (1) Compared with EL speech, the converted samples of the four systems have richer variations of spectral and F0 contours and present a duration that more closely aligned with the target, (2) Proposed 2 and 3 performed the best in converting the spectrogram shape and F0 to the target, which is consistent with their superior quantitative results in Table 3.5, and (3) Proposed 3 showed a slightly inappropriate pitch variation, reflecting the impact of lower-quality SD.

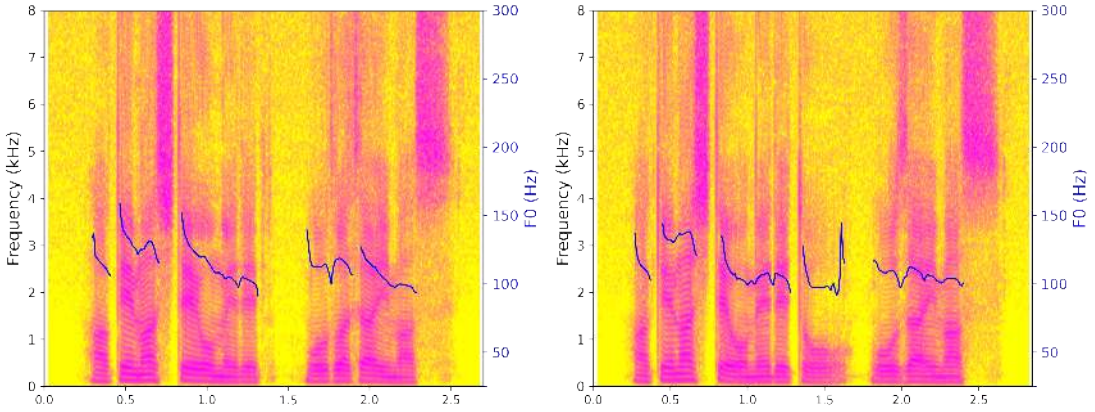
3.4.4 Subjective Evaluation Results

Subjective evaluations were conducted by randomly selecting outcomes from Proposed 1 and 2 using the datasize of 4,000. Figure 3.7 visualizes the obtained results, where seven types of speech were mixed into the tests for Cases 1 and 2. We can observe that both seq2seq baselines provide a reasonable naturalness ($\approx$ 2.7), even



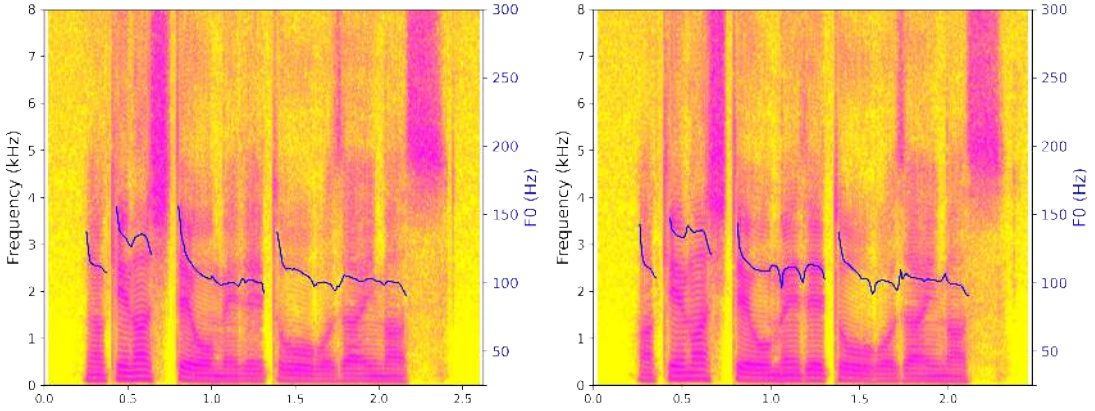
(a) Source EL.

(b) Target normal.



(c) Seq2seq baseline.

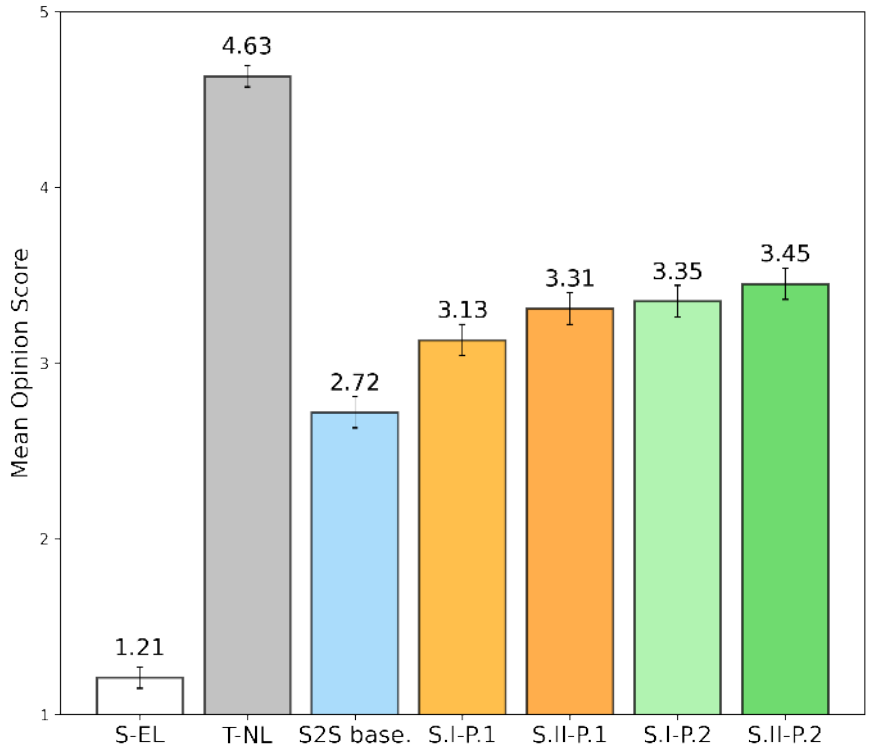
(d) Proposed 1.



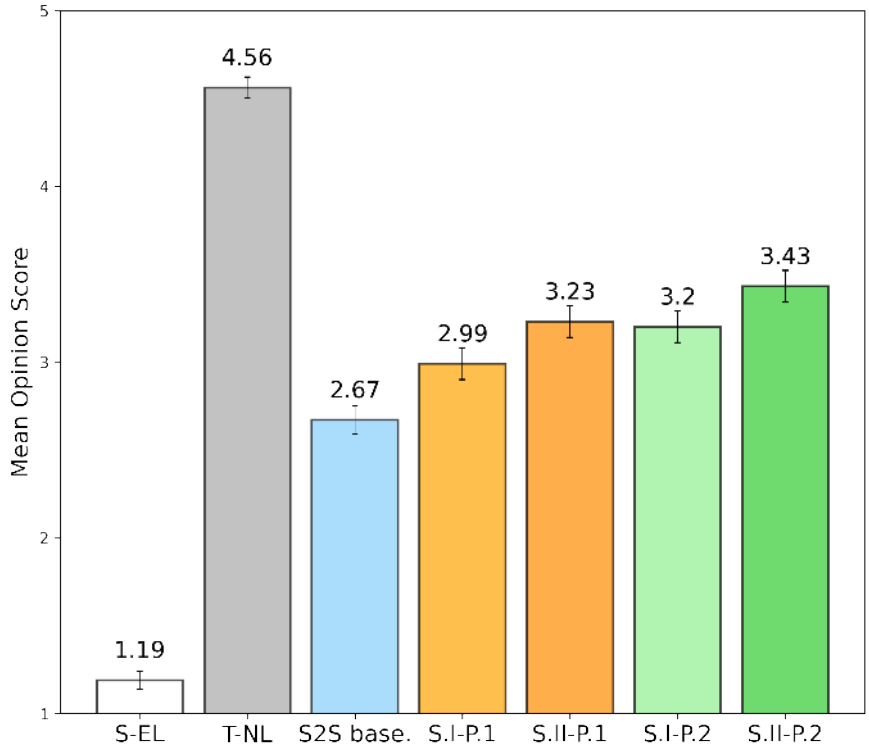
(c) Proposed 2.

(d) Proposed 3.

Figure 3.6: Spectrograms and F0 contour plots for a converted example 'a ke ma shi te o me de to u go za i ma su' using four systems in Case 2, where Proposed 1, 2, and 3 show the final (Stage II) output of 4,000 datasize. The horizontal axis displays time in seconds. The vertical axis presents spectral and F0 frequency.



(a) Case 1: SimuEL–NormSP.



(b) Case 2: RealEL–NormSP.

Figure 3.7: Mean Opinion Score (MOS) results with 95% confidence intervals for naturalness. The labels are P.1 and P.2 for Proposed 1 and 2, S.I and S.II for Stages I and II, S-EL and T-NL for source EL and target normal speech, respectively.

though the intelligibility is non-ideal. This shows that the listeners prioritized certain aspects such as fluency and voice quality over linguistic contents. Moreover, the proposed systems yielded significantly higher performance starting from Stage I than the baselines.

We can see that Proposed 2 outperformed Proposed 1 in both stages. Interestingly, in Case 1, Stage I of Proposed 2 already marginally surpassed Stage II of Proposed 1. In addition, the results of the proposed methods from Stages I to II reveal a consistent and stable improvement, which is similar to the findings from the objective results.

3.5 Conclusions

This chapter proposed novel training methods to enhance data efficiency and transferability of seq2seq VC on EL2SP via efficient pretraining and fine-tuning techniques, specifically based on low-quality SD. The concept of an innovative encoder adaptation training was introduced and designed by extending the typical pretrained seq2seq VC model with low-quality SD. Subsequently, still leveraging low-quality SD, a two-stage fine-tuning method was proposed that incorporated the updated components from the adaptation training, resulting in a more comprehensive framework.

Through a series of comparative experiments, it was demonstrated that even when using SD with limited fidelity, the proposed method could effectively improve EL2SP performance. Beyond achieving superior results compared to both non-seq2seq and seq2seq baselines under identical data constraints, various systems were developed by integrating adaptation training into the framework and redesigning specific training stages. Key determinants including training frameworks, original data, SD sizes and qualities, and representation discrepancies were systematically investigated.

The main observations are as follows: (1) Low-quality SD alleviated the data requirements and verified the robustness of the proposed method, whose performance of Stage I was positively correlated with the increase in SD quantity, (2) Proposed systems

that used encoder adaptation training as the intermediate step achieved the best performance, demonstrating that domain representations can be more smoothly adapted to the downstream EL2SP dataset through this approach, and (3) The two-stage fine-tuning scheme helped remedy the negative impact introduced by low-quality SD, progressively refining model performance and stabilizing the final outputs for EL2SP.

These results affirm the importance of architecture-aware transfer learning and highlight the practical advantages of the proposed training methods. Furthermore, the design decisions made at each stage of the proposed systems manifested different levels of effectiveness, underscoring their flexibility. Therefore, given current progress, we can expect that the proposed systems may offer some preliminary insights for other VC tasks faced with the problem of limited data, such as EL2SP under more severely interfered conditions, which will be described in the next chapter.

4 Robust Training Techniques for Electrolaryngeal Speech Enhancement in Noisy and Reverberant Conditions

This chapter proposes a novel training method for robust sequence-to-sequence (seq2seq) ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP), which addresses not only the constraints of low-resource original data but also the challenges posed by complex interference conditions. As noted in Subsection 1.3.2, due to the inherent complexity of EL2SP mapping, existing research on handling real-world interferences for EL2SP remains very limited. Unlike typical systems that rely on more complex network architectures or well-trained extra Speech Enhancement (SE) modules for robustness, the proposed method relies solely on a single seq2seq Voice Conversion (VC) framework using limited data. Inspired by the transfer learning strategies introduced in Chapter 3 for low-resource conditions and grounded in an understanding of seq2seq modeling mechanisms, this chapter proposes a unified two-stage fine-tuning strategy aimed at progressively improving model performance under complex real-world environments, including clean, noisy, reverberant, and mixed conditions. Within this framework, the low-quality parallel Synthetic Data (SD) generated from minimal original EL2SP data are leveraged here as well. Given the demonstrated po-

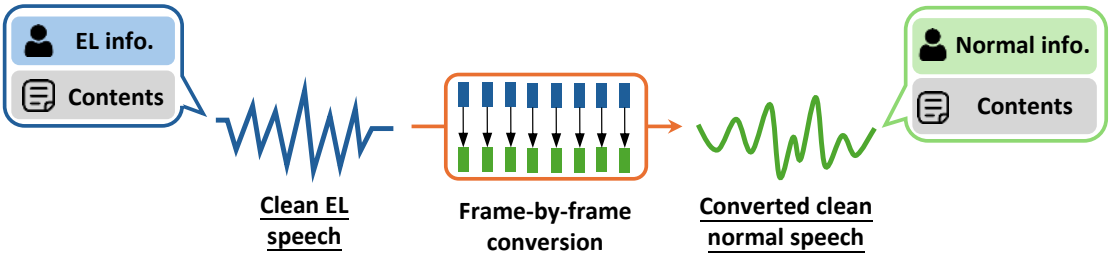
tential of SD in addressing low-resource challenges in Chapter 3, similar effectiveness is expected for this task. More crucially, by injecting diverse noise and reverberation —representing fine-grained real-world scenarios— into the EL SD, a many-to-one mapping is constructed with the target normal speech. This setup enables the model to simultaneously achieve robust adaptation to complex real-world scenarios as well as reasonable conversion.

Furthermore, driven by the data-centric nature of the training method, multiple systems are built with varying levels of fine-tuning and comprehensive evaluations are conducted in this chapter. First, the positive impact of EL SD containing more fine-grained real-world scenarios is verified on the final system performance. Then, the proposed systems are benchmarked against two types of baselines: one trained only on clean data, and another augmented with SE modules. These baselines reveal the extent to which traditional methods rely on SE components. Meanwhile, the proposed systems are shown to clearly outperform both categories, non-trivially handling clean and noisy-reverberant EL speech. Lastly, the intermediate representations of the proposed systems are analyzed to comprehend how they manage interferences, with visualization results further reflecting the effectiveness of the modeling. Altogether, these findings highlight the potential of seq2seq VC for interference-robust EL2SP and shed lights on promising future directions for improvement.

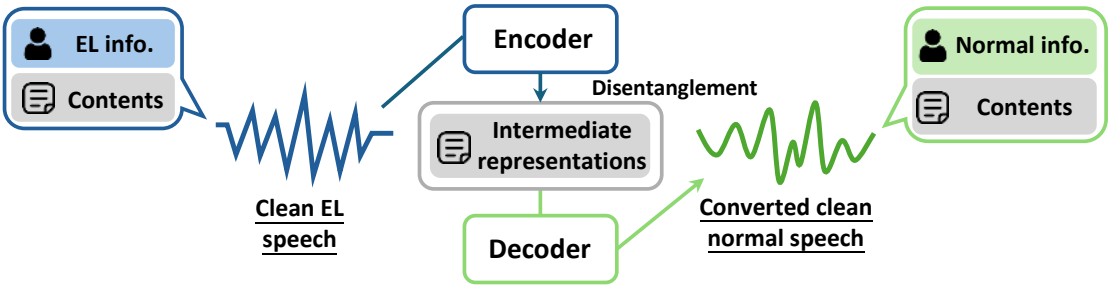
The organization of this chapter is as follows. Section 4.1 introduces the background, motivation, and main contributions. Section 4.2 presents the proposed many-to-one transfer learning method and comparable baselines. Section 4.3 describes the experimental evaluation settings, followed by comprehensive experimental results and analyses in Section 4.4. Section 4.5 concludes this chapter.

4.1 Introduction

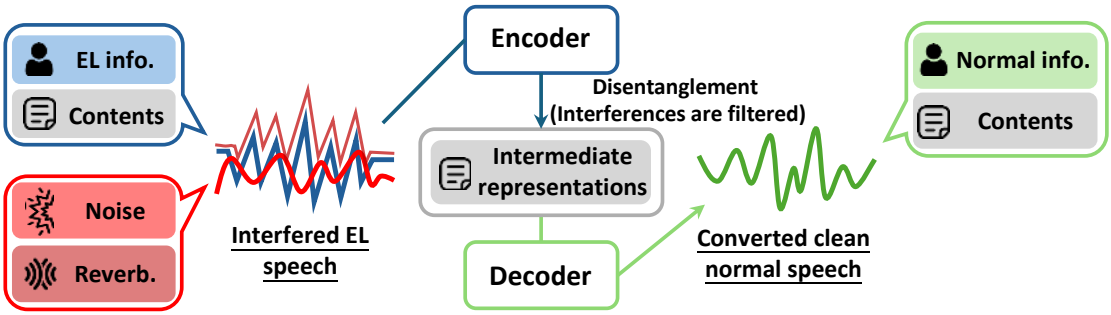
This section briefly introduces the motivation and contributions of the proposed method. As discussed in Subsections 1.2.1 and 1.2.2, seq2seq VC differs fundamentally



(a) Conventional VC based on frame-wise function.



(b) Sequence-to-sequence (Seq2seq) VC that disentangles the spoken content of clean EL speech.



(c) Seq2seq VC adapted to real-world EL2SP can generate intermediate representations by filtering out the interferences contained in EL speech.

Figure 4.1: Overview of Voice Conversion (VC) techniques for ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP).

from conventional frame-wise VC in its mapping mechanism, as illustrated in Figure 4.1(a) and (b). Frame-wise VC relies on explicit frame-to-frame alignment, which hinders accurate modeling of time-variant features. In contrast, seq2seq VC provides a different strategy by decomposing the conversion paradigm [147]; the encoder first disentangles linguistic contents from the source speech features as intermediate representations, and the decoder reconstructs the converted speech based on these representations and the target features, as demonstrated in Subsection 2.1.2. This allows the seq2seq model to flexibly determine the output length, making it well-suited for

EL2SP tasks, in which the strength has already been demonstrated in Chapter 3.

Furthermore, the training techniques introduced in Chapter 3 effectively enhance the performance of seq2seq models under limited EL2SP datasets. Nevertheless, most prior EL2SP studies focus exclusively on high-quality recordings in simple, clean environments, aiming to isolate and resolve the core mapping challenges between EL and normal speech, as emphasized in Subsection 1.3.2. However, real-world scenarios often contain various interferences, such as noise and reverberation, making robustness under such conditions an essential research objective.

As reviewed in Section 2.2, many existing methods to interference robustness rely on complex frameworks, including external SE modules or extended representation learning networks. Although effective, this approach increases the inference complexity and still depends on large-scale clean corpora, which is an impractical assumption for real-world EL2SP applications. Moreover, most are tailored to specific types of interference and struggle under compound conditions (e.g., overlapping noise and reverberation), with their applicability to EL2SP yet to be validated. Therefore, regarding the perspective of architectural simplicity and generalizability to real-world interferences, the practicality of such an approach remains insufficient, thereby calling for more efficient alternatives.

Considering the fundamental deficiencies of these sophisticated frameworks for real-world EL2SP, the proposed method builds on the successes of training techniques for addressing low-resource data in Chapter 3, especially for the data augmentation and fine-tuning methods used. Starting from a new perspective of adaptation, the proposed method aims to develop a noise- and reverberation-robust seq2seq EL2SP system using small-scale clean data, which can realize a direct conversion for interfered EL speech. Specifically, the seq2seq architecture is maintained and EL speech is input with diverse noise and reverberation properties to fine-tune the model for fine-grained real-world scenarios. An important motivation here is derived from the definition of VC, i.e., the objective of an effective encoding is always pure linguistic

intermediate representations closely tied to speech information. When VC adapts to real-world scenarios, we can expect to treat non-speech interferences as extraneous information and filtered out, as depicted in Figure 4.1(c). Another motivation is that the conversion target in this study is always clean normal speech. This clear objective not only aligns with major settings in real-world EL2SP, but also makes the decoder contribute to improving the accuracy of intermediate representations by optimizing towards a well-defined goal. Note that the training data are augmented by simulating EL SD with noise and reverberation properties to facilitate addressing real-world EL2SP tasks. Given the benefits of low-quality EL SD for clean EL2SP in Chpater 3, incorporating such imperfect data with interfered information is presumed to somewhat enrich knowledge and disentangle speech and non-speech information, leading to better conversion performance.

The work in this chapter aims to improve transfer learning for real-world seq2seq EL2SP by developing training methods that utilize data augmentations with different attributes. With the stepping stone provided by preliminary studies [45], [78], this chapter not only designs various systems in pursuit of optimized performance, but also presents a comparative study through systematic experiments. The contributions are summarized as follows:

- This work represents a new effort to cope with real-world EL2SP using practically limited clean data. In particular, a many-to-one framework is proposed that integrates different noisy and reverberant EL SD with corresponding clean normal SD, enabling a simultaneous handling of multiple interferences without requiring any extra module, label, or strict alignment pattern.
- Four systems of different fine-tuning levels are designed on the basis of various EL data. Moreover, following the study outlined by Choi *et al.* [70], both denoising and dereverberation SE modules are introduced to the EL2SP systems. Apart from using SE modules pretrained on normal datasets, they are further fine-tuned on EL data through either cascading or joint training approaches,

ensuring more optimized EL2SP baselines. Nonetheless, the best-performing proposed system outperforms all these baselines.

- The two-stage fine-tuning strategy, initially proposed for the seq2seq-based EL2SP systems in Chapter 3, is extended to SE network training. Experimental results confirm that this effectively enhances the robustness of SE modules, offering benefits for downstream EL2SP and further demonstrating methodological generalizability of the two-stage fine-tuning.
- For the first time, the hidden representation spaces of learned EL2SP systems are visualized when dealing with real-world conditions and how those are related to the performance.
- Through objective and subjective experiments, the proposed systems under conditions of clean-, noisy-, and/or reverberant-EL speech are evaluated. The reasonable results obtained in terms of speech quality, naturalness, and speaker similarity verify their generalizability.

4.2 Many-to-one Transfer Learning Method

Recalling the main challenge of enhancing the robustness of seq2seq EL2SP in the presence of acoustic interferences, especially when only a limited amount of original EL2SP data is available, this chapter focuses on combining easy-to-obtain Synthetic Data (SD), particularly the simulated EL SD with varying acoustic properties, with transfer learning to improve transfer performance in real-world environments. Figure 4.2 illustrates the process of developing the proposed method, which consists of two main parts: Part 1: Data augmentation (on the left) and Part 2: EL2SP training (on the right).

Starting with Part 1, two Text-To-Speech (TTS) models for EL and normal speech are fine-tuned using the original EL2SP dataset. Then the two models are used to generate parallel SD, which contains EL and normal SD. Here, EL SD and the original

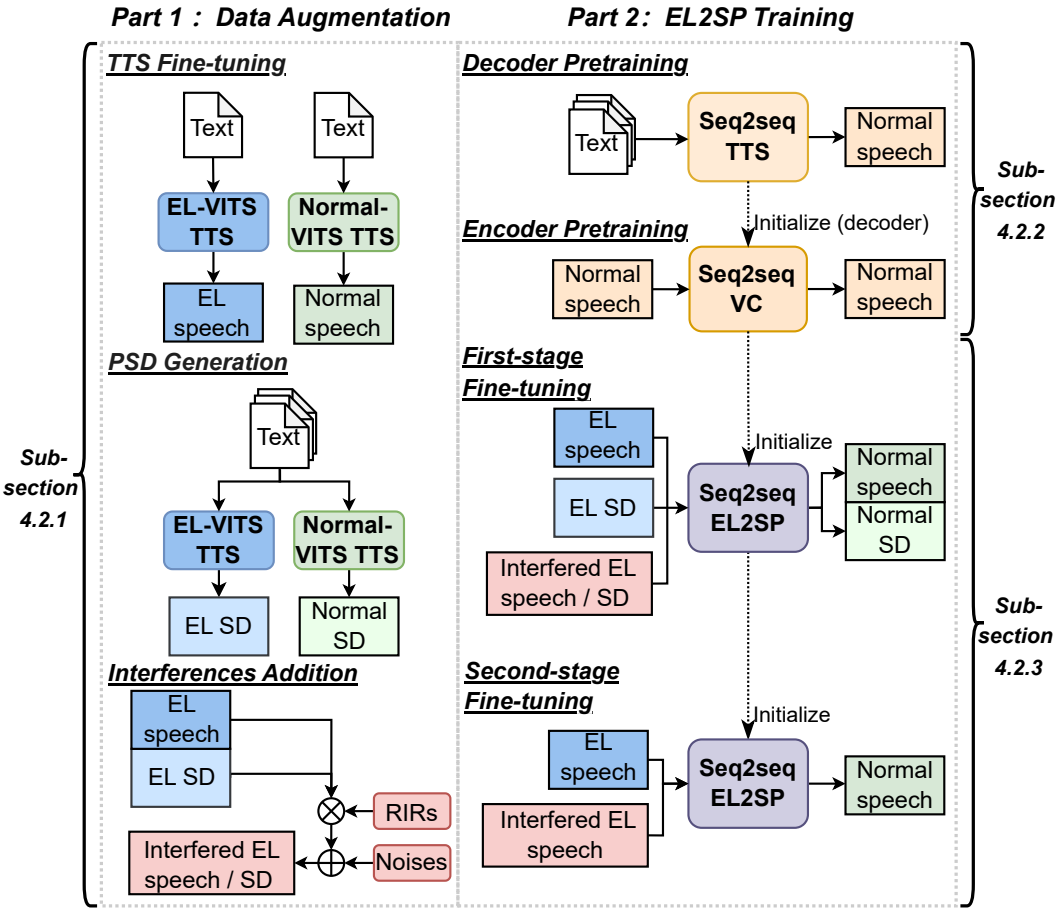


Figure 4.2: Overview of the proposed method for building real-world EL2SP, with the subsections labeled to correspond to the description of each process.

EL data are further *interfered* by adding noise, reverberation, or a combination of both (Subsection 4.2.1).

Moving to Part 2, the pretraining–fine-tuning stages are carried out. First, a pretrained seq2seq VC model is developed by TTS pretraining on a normal TTS database (Detailed in Subsection 4.2.2). This is then followed by a two-stage fine-tuning by incorporating the data augmentations to achieve the final EL2SP system, which enhances robustness under noisy and reverberant conditions (Subsection 4.2.3).

Lastly, multiple systems are constructed by adjusting the types of interfered data used during the fine-tuning process. Moreover, several typical baseline architectures

are designed by leveraging extra SE modules for fair comparisons with the proposed method (Subsection 4.2.4).

4.2.1 ElectroLaryngeal (EL) and Normal Text-To-Speech (TTS) Fine-tuning for Parallel Synthetic Data (SD) Generation

Given that only small-scale original data for EL2SP is available in this work, the process in this subsection, resembling that in Subsection 3.2.2, aims to augment the training data by producing large-scale parallel SD. But a pretrained Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS)-based model [148] is separately fine-tuned via original source EL and target normal speech sets, for faster turnaround than the previous setup in Chapter 3. Thus the corresponding VITS-based EL-TTS and normal-TTS models are obtained, respectively. After this, the same external text set is input into both EL-TTS and normal-TTS models to generate parallel SD. Note that, owing to the low-resource dataset for TTS fine-tuning, the quality of parallel SD is still poor.

Once the parallel SD is generated, a much larger EL2SP dataset is obtained, in which both EL speech and EL SD are further injected with the unique interferences including different types of background noise and/or reverberation, with each EL utterance corresponding to a specific interference. It is worth pointing out that most interferences are leveraged by the EL SD, owing to its size being much larger than the original EL data. All these interfered and clean EL datasets are then used in the subsequent fine-tuning stages.

4.2.2 Pretraining of Sequence-to-sequence (Seq2seq) Voice Conversion (VC) Model

Inspired by the work in Chapter 3, this chapter also adopts Voice Transformer Network (VTN) [82] to build a pretrained *one-to-one* seq2seq VC model by utiliz-

ing Transformer-based TTS pretraining, whose generated pure linguistic intermediate representations remain essential for achieving effective seq2seq EL2SP under real-world interferences. Motivated by the same decoding mechanism between TTS and VC, the proposed method aims to transfer the compact, rich-linguistic representations, derived from a normal TTS corpus through attention mechanism, while also sharing the speech decoder from pretrained TTS onto VC. A significant advantage of this approach is that it only requires an arbitrary single-speaker corpus and the corresponding transcriptions to acquire pretrained knowledge, rather than necessitating parallel corpora of the same magnitude. This largely relaxes the constraints for developing seq2seq VC model.

As introduced in Subsection 3.2.1, VTN pretraining includes the decoder and encoder pretraining. Initially, decoder pretraining involves a typical TTS training with a large normal TTS dataset, which enables the decoder to effectively associate speech features with corresponding pure linguistic information from the encoded text. Following this, during the encoder pretraining, the TTS corpus serves as both input and target. A new speech encoder, following an autoencoder training style, is then updated using a reconstruction loss by keeping the parameters of the pretrained decoder fixed. In this manner, the encoder is forced to learn to extract the rich-linguistic representations from the speech signals instead of from the text, owing to the inherited intermediate representations and retained ability of the fixed decoder to recognize linguistic information.

4.2.3 Two-stage Many-to-one ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP) Fine-tuning

At the beginning of the fine-tuning process, the pretrained VTN is employed to impart the effective *a priori* parameters for initializing the seq2seq EL2SP model, ensuring improved transferability and more efficient convergence speed compared with

training from scratch. A large-scale EL2SP dataset is constructed for the first-stage fine-tuning, where the original EL data, EL SD, and interfered EL data are pooled together as inputs. Hence, the original target normal data and normal SD are repeatedly used to form the parallel pairs with their corresponding EL inputs. These training pairs are fed into the model for training in a many-to-one mapping manner, i.e., mapping interfered and clean EL speech to clean normal speech, to provide the vast knowledge for generalized performance. However, some distorted properties contained in SD might negatively affect the accuracy of model weights. Therefore, in the second-stage fine-tuning, aimed at further refining the model parameters, clean and noisy-reverberant versions of original EL data are used, paired with their corresponding normal data, to finalize the EL2SP model.

At the core of this method is the concept of learning stronger perception from various acoustic properties in different types of EL speech. Thus, using more subdivided properties, i.e., clean, noisy-only, reverberant-only, and noisy-reverberant EL inputs, would facilitate the adaptation to real-world scenarios. This naturally motivates us to design different systems, which will be introduced in Subsection 4.2.4.

4.2.4 Proposed and Baseline Systems

4.2.4.1 Proposed Systems

As depicted in Table 4.1, four systems named *Proposed 1*, *Proposed 2*, *Proposed 3*, and *Proposed 4* are designed. These systems share the same seq2seq framework, whereas the main difference between them is EL inputs with different acoustic conditions used during fine-tuning stages as follows:

- *Proposed 1*: Since no SD is used, Proposed 1 can be viewed as a standard approach for adapting to real-world scenarios by conducting a direct fine-tuning on the original EL data containing clean/noisy-reverberant conditions.
- *Proposed 2*: Proposed 2 undergoes the two-stage fine-tuning process. In the first-stage fine-tuning, clean and noisy-reverberant original/synthetic EL inputs are utilized, whereas in the second-stage fine-tuning, the same training data as in

Table 4.1: Types of input EL training data for individual systems. Here, "ORG" and "SYN" indicate whether the EL speech is the original or synthetic, while "C", "R", "N", and "NR" represent the following conditions: Clean, Reverberant, Noisy, and Noisy-Reverberant, respectively.

Systems	First stage		Second stage	
	ORG/SYN	Conditions	ORG/SYN	Conditions
Proposed 1	Yes/No	C + NR	—	—
Proposed 2	Yes/Yes	C + NR	Yes/No	C + NR
Proposed 3	Yes/Yes	C + NR + N + R	Yes/No	C + NR
Proposed 4	Yes/Yes	C + NR + N + R	Yes/No	C + NR + N + R

Proposed 1 is applied.

- *Proposed 3*: Compared with Proposed 2, Proposed 3 performs the same second-stage fine-tuning, but the types of EL data used in the first-stage fine-tuning are expanded to further incorporate inputs with only noise or reverberation.
- *Proposed 4*: The types of EL data used in the second-stage fine-tuning may also impact the training performance. Therefore, in Proposed 4, on the basis of the consistent first-stage fine-tuning with Proposed 3, the second-stage fine-tuning differs slightly by additionally leveraging the original EL inputs with only noise or reverberation.

4.2.4.2 Baseline Systems

To figure out various properties of the proposed method in terms of interference-robust EL2SP, a comparative study is conducted by introducing several baseline systems.

1) Baseline systems adapted to clean environment: As the first question, we shall look simply at how well the proposed systems perform compared with those adapted solely to a clean environment. To contextualize this, two fundamentally off-the-shelf baseline systems are prepared, namely *Baseline 1* and *Baseline 2*, which fine-tune the pretrained VTN that is identical to the proposed systems, but without using any form of interfered data. Baseline 1 uses only the original clean pairs, whereas Baseline 2

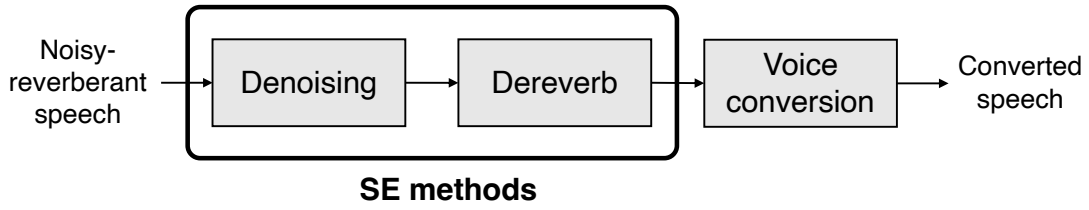
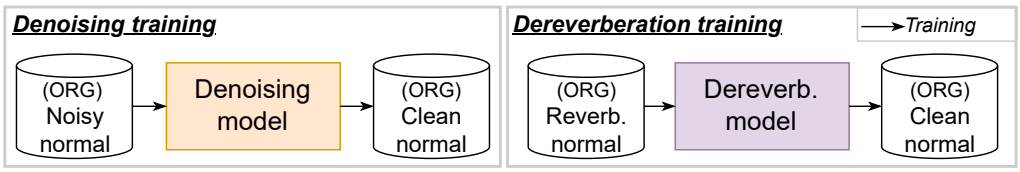


Figure 4.3: Overview of the baseline frameworks using Speech Enhancement (SE) methods.

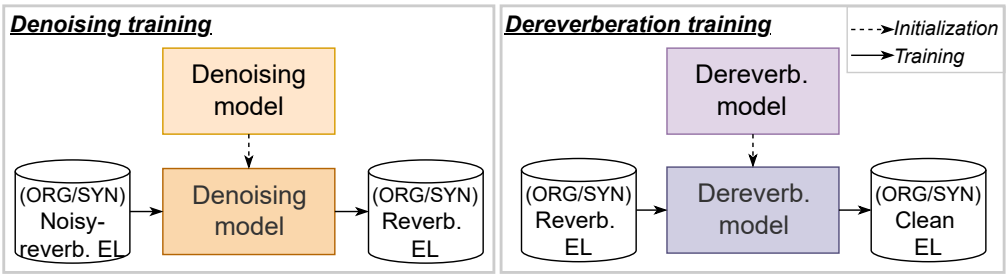
additionally uses the same low-quality parallel SD as that of Proposed 2, 3, and 4.

2) Baselines using SE methods: Another question we shall consider is how effective the proposed systems are under noisy-reverberant condition from an architectural aspect, especially compared with mainstream systems equipped with SE modules. As a result, aside from the comparisons for the proposed systems with Baselines 1 and 2, four SE methods that are connected to the same EL2SP framework are designed, to establish corresponding baseline systems. Note that these baselines, following the enhancement order demonstrated to achieve state-of-the-art performance by Choi *et al.* [70], first conduct denoising and then dereverberation for noisy-reverberant inputs, as shown in Figure 4.3. The SE modules, named *Extension-pretrain (E-pt)*, *Extension-fine-tuning (E-ft)*, *Extension-ft-cascade (E-ft-c)*, *Extension-ft-joint (E-ft-j)*, and *Extension-two-stage-ft-joint (E-2ft-j)* are mainly distinguished according to their specific training methods, which are summarized in Figure 4.4.

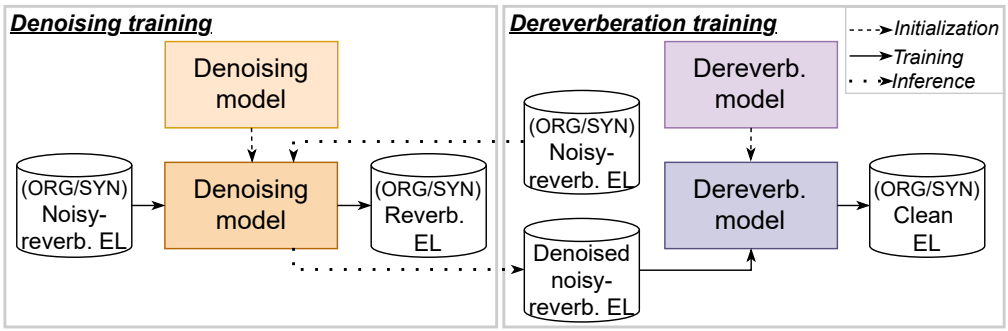
- *E-pt*: As illustrated in Figure 4.4(a), E-pt provides a common SE strategy that involves using interfered data from widely available original normal human corpora to train denoising and dereverberation models. These SE models are then directly used to enhance interfered EL speech.
- *E-ft*: Owing to significant differences in acoustic properties between EL and normal speech, the direct use of the two SE modules generated by E-pt may not generalize well to EL speech. Thus, the interfered EL speech, simulated using both the original and synthetic clean EL speech datasets, is further used as downstream data to separately fine-tune the two SE modules, thus constructing E-ft, as depicted in Figure 4.4(b).



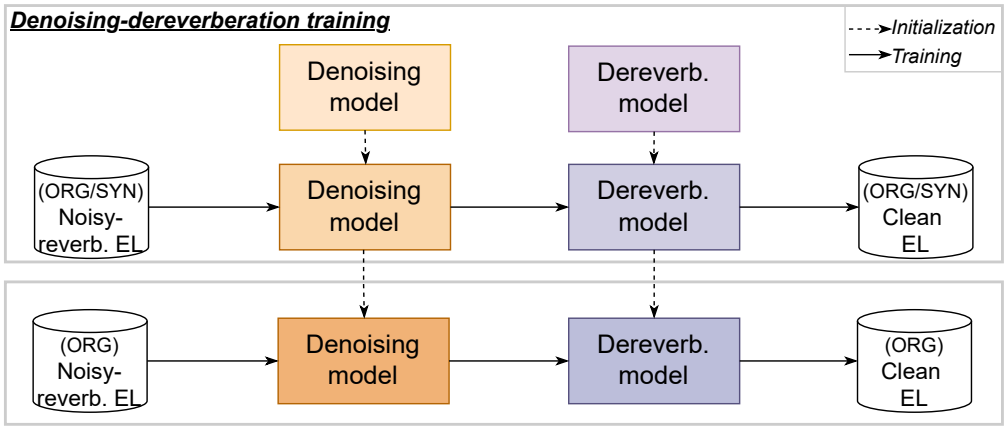
(a) Extension-pretrain (E-pt)



(b) Extension-fine-tuning (E-ft)



(c) Extension-ft-cascade (E-ft-c)



(d) Upper: Extension-ft-joint (E-ft-j). Lower: Extension-two-stage-ft-joint (E-2ft-j)

Figure 4.4: Training methods for SE models that are connected to EL2SP, where E-ft, E-ft-c, and E-ft-j are all initialized by the pretrained modules from E-pt. E-2ft-j follows the same joint framework as E-ft-j but undergoes two-stage fine-tuning, with an additional initialization using the parameters of E-ft-j. ORG and SYN indicate whether the training data are original or synthetic.

- *E-ft-c*: As plotted in Figure 4.4(c), E-ft-c follows the fine-tuning using the interfered EL speech of the same volume as E-ft, but enhances the process by integrating the denoising and dereverberation models in a cascade, where the output from the denoising model serves as input for training the dereverberation model.

- *E-ft-j*: As indicated in Subsection 2.2.2, intermediate output would inevitably bring some distortions, potentially limiting SE performance. Therefore, E-ft-j, illustrated in the upper block of Figure 4.4(d), while inheriting the pretrained weights from E-pt, combines the two SE modules into a joint network, utilizing both original and synthetic data for training. This method ensures that the training data only includes noisy-reverberant and clean EL data, without intermediate generation during SE inference.

- *E-2ft-j*: Motivated by the two-stage fine-tuning methodology in the proposed seq2seq VC framework, here, it is hypothesized that this approach can also be generalized to SE training. To demonstrate its generalizability, E-ft-j is further extended. Specifically, the same joint framework and initial training process as E-ft-j are adopted for the first fine-tuning stage, where the pretrained SE model is fine-tuned using interfered synthetic and original EL data, along with their clean counterparts. In the second stage, the interfered and clean original data are solely used to refine the final SE model, thus establishing E-2ft-j, as illustrated in the lower block of Figure 4.4(d).

These SE methods are then individually ensembled with various EL2SP systems to deploy the corresponding baseline systems with higher comparative appeal. Three types of baseline systems are specifically designed based on these SE methods, all of which adopt the same inference process: The SE modules first process the interfered EL inputs to generate enhanced data, which is subsequently converted by the downstream EL2SP part.

- The first type of system employs a straightforward approach, directly combining

the SE module with an EL2SP model trained on clean data. On the basis of this approach, four systems are formed by pairing E-pt, E-ft, E-ft-c, and E-ft-j with Baseline 1, resulting in systems named *E-pt-Base1*, *E-ft-Base1*, *E-ft-c-Base1*, and *E-ft-j-Base1*, respectively. Although some distortions are contained in processed EL data, these systems are expected to achieve better conversion results than Baseline 1. They are evaluated to determine (1) how the distortions affect the converted speech when downstream model is trained only on clean data, and (2) the performance differences with other proposed systems.

- The second type aims to further improve the adaptability of the downstream EL2SP to the processed EL data. Here, E-ft-c and E-ft-j are specifically used to separately process the interfered version of the training data from Baseline 1. The processed data are then used to fine-tune their respective downstream EL2SP models. Accordingly, these two systems are named *E-ft-c-d-Base1* and *E-ft-j-d-Base1*.

- The last one follows a similar manner to the second one while processing the much larger-scale, interfered version of the training data from Baseline 2, which is used to investigate the potentially positive impact of low-quality SD. For this, E-ft-j and E-2ft-j are utilized as the SE modules and the entire systems are named *E-ft-j-d-Base2* and *E-2ft-j-d-Base2*, respectively.

Therefore, progressively deeper baselines are constructed, taking into account the factors including SE performance, adaption to processed EL data, and SD effects, to facilitate systematic comparative studies between them and the proposed systems.

4.3 Experimental Evaluation Settings

This section first outlines the experimental protocol (Subsection 4.3.1), which comprises the datasets used, model architectures with their implementation configura-

tions, followed by the metrics (Subsection 4.3.2) for evaluating the experimental results including objective and subjective assessments.

4.3.1 Experimental Protocol

In this subsection, an overview of the experimental protocol is presented, including the dataset setup and model configurations.

4.3.1.1 Datasets

Three types of datasets used for the experiment are as follows:

- *TTS database*: To accomplish the pretraining mentioned in Subsections 4.2.1 and 4.2.2 for the seq2seq VC and VITS TTS models, all 7,696 utterances in the Japanese JSUT database [141] were utilized, which amounts to approximately 10 hours of speech. All transcriptions from the JSUT database were also selected for generating the parallel SD.

- *Original EL2SP datasets*: To develop and evaluate EL2SP systems, two small-scale and semi-parallel EL2SP datasets were constructed, referred to as Patient 1 dataset and Patient 2 dataset, respectively, with totally different utterance contents. Both were recorded under identical recording conditions; in a professional soundproof booth, using a Shure SM58 dynamic microphone, and a Roland Rubix 22 audio interface connected to the Audacity recording software[1]. All speakers involved in the recordings were native Japanese speakers.

  - *Patient 1 dataset:* It consists of 200 EL utterances totaling less than 10 minutes, and 413 normal utterances around 20 minutes. The EL speech was recorded from a male laryngectomee using an electrolarynx. Due to a complete laryngectomy for cervical esophageal cancer, his normal (pre-surgery) speech was unavailable. To provide a reference for healthy speech,

[1] https://www.audacityteam.org/ (Version 3.1.2, Accessed: Dec. 4, 2021)

a healthy male speaker recorded the normal speech under the same recording conditions.

  - *Patient 2 dataset:* It includes 573 EL utterances totaling approximately 29 minutes, and 373 normal utterances totaling roughly 18 minutes. This dataset was recorded from a male laryngectomee diagnosed with severe hypopharyngeal cancer. His larynx was completely removed, and he underwent a jejunal graft transplantation from his abdomen. His normal speech was recorded prior to the surgery. Despite the presence of the disease, his vocal cords were not significantly affected at the time of recording, so his normal voice remained largely unaffected. Given these conditions, this dataset simulates a scenario where pre-surgical speech from laryngectomees is available.

To address the semi-parallel nature of Patient 1 and Patient 2 datasets, the fine-tuned TTS models were used to add the corresponding EL SD (213) for the Patient 1 dataset and normal SD (200) for the Patient 2 dataset, respectively, maximizing the utilization of all feasible original data during EL2SP training. In both datasets, the development and test sets consisted of 20 and 40 original clean utterances, along with their noisy-reverberant counterparts, while the remainder was used for training.

- *Noise and reverberation settings*: 8,109 and 8,269 noise clips along with their corresponding Room Impulse Responses (RIRs) were leveraged from the WHAMR! dataset [149], to generate interfered EL data for Patient 1 and Patient 2 datasets, respectively. Figure 4.5 depicts the overall process of simulating interfered EL speech based on the clean EL speech. First, the reverberant versions of the original and synthetic EL data were created by convolving them with RIRs, whose reverberation time (T60) range was from 0.1 to 1.0 seconds. Subsequently, noise clips with five Signal-to-Noise Ratios (SNRs) of 0, 5, 10, 15, and 20 dB were mixed

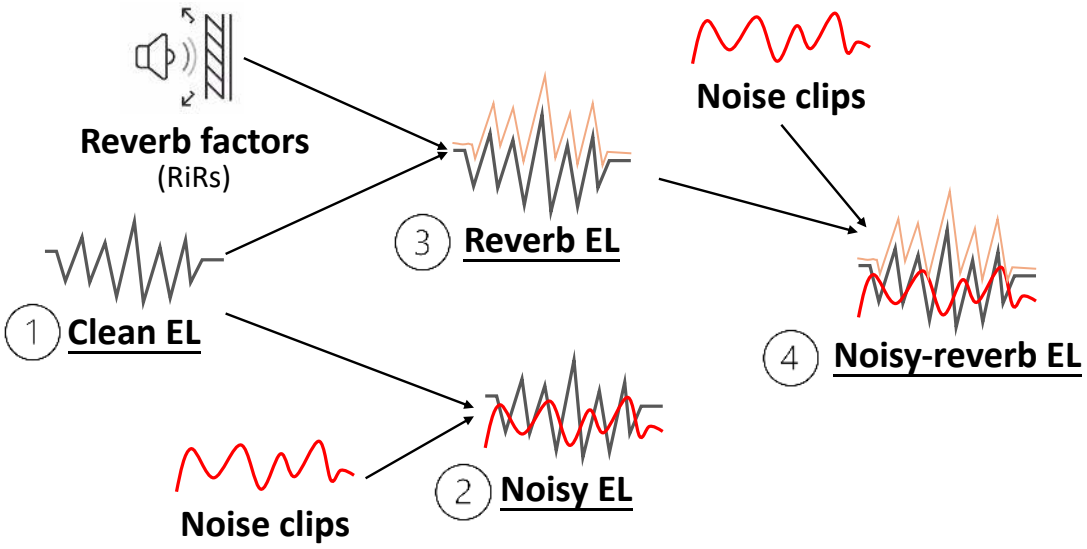


Figure 4.5: Illustration of interference addition using all available clean EL data, including original and synthetic. Apart from the clean condition, three types of interfered conditions are sequentially simulated: noisy, reverberant, and noisy-reverberant.

with all the clean and corresponding reverberant EL data to create noisy and noisy-reverberant versions. Each EL utterance was assigned a unique noise clip and a distinct set of RIR parameters, thereby ensuring both the diversity of the dataset and the complete separation between training and test sets. All noise clips and RIR configurations in the training set, including those applied to the original EL data and EL SD, were strictly excluded from those in the test set. Thus, all interferences present in the test set were entirely unseen during model training.

Given the necessity for SE methods in establishing baselines, two datasets were additionally introduced for training the denoising and dereverberation models of E-pt. Following the work presented by Fujimura and Toda [150], 1,000 utterances from the LibriSpeech dataset [151] were used here, which were mixed with noise clips of the CHIME3 dataset [152], to conduct denoising training. Concurrently, still using LibriSpeech as the basis, 10,000 clean speech samples were dynamically converted into reverberant speech on-the-fly to adequately pretrain the dereverberation model. Afterwards, the interfered original/synthetic EL data were further employed as the

fine-tuning dataset to establish E-ft, E-ft-c, E-ft-j, and E-2ft-j.

4.3.1.2 Configuration Settings

The EL data were initially processed at 16 kHz during interference simulation for the training of denoising and dereverberation modules. During EL2SP training, both EL and normal data were then resampled at 24 kHz and processed using the 80-dimensional Mel filterbanks with 2,048 FFT points and a 300-point shift to extract acoustic features. The Transformer-based pretrained seq2seq VC and VITS TTS models accomplished with the ESPnet toolkit [96], [142], [153], similar to the implementation in Chapter 3, and followed the official configurations.

Additionally, denoising training was completed using a complex Time-Frequency Mask (TFMask) network [154] and specifically referred to the configurations by Fujimura and Toda [150]. Leveraging the Asteroid platform [155], the dereverberation module applied Conv-TasNet [156] as the backbone, following its original configurations. Parallel WaveGAN (PWG) neural vocoders [144] were used to reconstruct the waveforms of the EL2SP outputs, while the parallel SD was synthesized directly from the VITS TTS models. Two speaker-dependent PWGs were trained from scratch, each corresponding to the target normal speech of Patient 1 and Patient 2 datasets, respectively.

4.3.2 Evaluation Metrics

4.3.2.1 Objective Evaluation

Following the same setup as in Chapter 3, two objective metrics were employed: (1) Mel-Cepstrum Distortion (MCD, in dB), which measures spectral quality by measuring distortions between target and converted speech, and (2) Character Error Rate (CER, in %), which reflects intelligibility of converted samples through an ASR engine trained as in [146], with their definitions provided in Subsection 2.4.1.

This experiment also explored the performance of the baseline systems that use extra SE methods. Hence, the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR, in

dB) and Short-Time Objective Intelligibility (STOI), as described in Subsection 2.4.1, were applied to evaluate the distortion and intelligibility of processed speech within the SE modules.

4.3.2.2 Subjective Evaluation

Following the protocols described in Subsection 2.4.2, Mean Opinion Score (MOS) and speaker SIMilarity (SIM) tests were carried out to evaluate the perceptual performance of the EL2SP systems. Six randomly selected converted samples of clean and noisy-reverberant EL input from Proposed 1, 2, and 3, and Baseline 1 were presented to each listener. Fifteen Japanese native speakers were recruited. Audio samples are available online[2].

4.4 Experimental Results and Analysis

This section reports a series of comprehensive objective evaluations and subjective listening tests to systematically present and analyze the proposed systems, comparing them with various baseline systems under specific real-world conditions. The investigated aspects include the effectiveness of the proposed two-stage many-to-one fine-tuning, performance discrepancies under different conditions, and an analysis of the intermediate representations in the proposed systems.

4.4.1 Objective Evaluation Results

Comparison with seq2seq baselines: All the proposed systems were compared with the baseline systems trained on clean data, during which different acoustic properties were simulated on the EL test set to obtain conversion results under various interfered conditions. The results for the Patient 1 dataset are documented in Tables 4.2, 4.3, and 4.4. Additionally, experiments on the Patient 2 dataset were conducted, using

[2] https://silenticymoon.github.io/APSIPA-demo/

Table 4.2: Objective evaluation results based on the Patient 1 dataset, where the inputs are Clean, Noisy (N), Reverberant (R), and Noisy-Reverberant (NR) EL conditions. Stage I and Stage II represent the first- and second-stage fine-tuning conducted for Proposed 2, 3, and 4, respectively.

Systems		**Clean EL**	**NR-EL**	**N-EL**	**R-EL**
		MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]
Proposed 1	—	6.44 / 39.2	7.11 / 55.9	7.08 / 54.3	6.82 / 42.3
Proposed 2	Stage I	5.89 / 29.3	6.33 / 38.4	6.00 / 33.0	6.06 / 30.9
	Stage II	5.71 / 27.0	6.09 / 36.7	5.83 / 32.5	5.88 / 27.1
Proposed 3	Stage I	5.80 / 27.8	6.30 / 37.4	6.10 / 33.3	5.93 / 30.2
	Stage II	5.69 / 24.5	6.06 / 35.9	5.89 / 33.0	5.78 / 27.0
Proposed 4	Stage II	5.61 / 25.3	6.02 / 34.6	5.82 / 32.8	5.74 / 27.9
Baseline 1	—	6.71 / 41.4	8.81 / 77.7	8.19 / 64.6	7.66 / 63.0
Baseline 2	—	6.15 / 32.1	11.29 / 86.9	10.11 / 69.1	9.60 / 70.5

Table 4.3: Objective evaluation results based on the Patient 1 dataset, when the EL input is noisy-reverberant with fixed Signal-to-Noise Ratios (SNRs) of −5, 2, 12, and 22 dB.

Systems		**SNR: −5**	**SNR: 2**	**SNR: 12**	**SNR: 22**
		MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]
Proposed 1	—	8.70 / 73.2	7.59 / 63.7	6.82 / 46.5	6.63 / 47.1
Proposed 2	Stage I	7.65 / 64.1	6.35 / 39.2	6.06 / 33.0	6.03 / 31.1
	Stage II	7.56 / 58.4	6.24 / 39.4	6.00 / 32.9	5.90 / 29.2
Proposed 3	Stage I	7.58 / 58.3	6.41 / 38.2	6.05 / 30.6	6.03 / 29.9
	Stage II	7.43 / 56.9	6.20 / 39.1	5.87 / 30.5	5.79 / 27.7
Proposed 4	Stage II	7.42 / 56.4	6.15 / 37.8	5.80 / 31.5	5.75 / 27.8
Baseline 1	—	11.04 / 74.4	10.29 / 75.8	8.76 / 83.9	7.98 / 63.7
Baseline 2	—	11.43 / 87.1	12.05 / 89.6	11.25 / 84.8	9.94 / 81.5

the same set of representative interfered test conditions as in Table 4.2. The corresponding results are presented in Table 4.5. Proposed 2 and 3 include results from both the first- and second-stage fine-tuning, whereas only the second-stage fine-tuning results are shown for Proposed 4. Note that Proposed 3 and 4 share the same first-stage

Table 4.4: Objective evaluation results based on the Patient 1 dataset, when the EL input is noisy-reverberant with the fixed T60s (1.0, 0.80, 0.40, and 0.20 seconds).

Systems		**T60: 1.0**	**T60: 0.80**	**T60: 0.40**	**T60: 0.20**
		MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]
Proposed 1	—	7.79 / 62.4	7.22 / 56.6	7.30 / 53.0	6.97 / 52.6
Proposed 2	Stage I	6.75 / 47.7	6.15 / 38.5	6.05 / 35.7	5.99 / 30.8
	Stage II	6.75 / 45.2	6.07 / 36.6	5.94 / 34.3	5.81 / 30.6
Proposed 3	Stage I	6.59 / 46.3	6.10 / 35.3	6.18 / 32.9	5.98 / 29.5
	Stage II	6.45 / 45.0	6.04 / 34.2	5.93 / 35.5	5.76 / 30.3
Proposed 4	Stage II	6.37 / 46.2	6.01 / 34.8	5.92 / 35.4	5.74 / 29.2
Baseline 1	—	9.71 / 73.6	9.52 / 78.5	8.63 / 66.0	8.03 / 61.3
Baseline 2	—	12.00 / 84.3	11.45 / 83.9	11.17 / 81.7	9.91 / 71.0

Table 4.5: Objective evaluation results based on the Patient 2 dataset, where the test inputs are Clean, Noisy (N), Reverberant (R), and Noisy-Reverberant (NR).

Systems		**Clean EL**	**NR-EL**	**N-EL**	**R-EL**
		MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]	MCD [dB] / CER [%]
Proposed 1	–	7.36 / 47.1	7.77 / 54.0	7.71 / 52.7	7.68 / 52.4
Proposed 2	Stage I	6.19 / 32.2	6.60 / 40.0	6.51 / 37.7	6.41 / 36.9
	Stage II	6.14 / 31.3	6.50 / 38.0	6.28 / 35.0	6.25 / 37.1
Proposed 3	Stage I	6.09 / 30.9	6.56 / 38.7	6.25 / 36.1	6.30 / 33.7
	Stage II	6.06 / 30.8	6.51 / 36.7	6.25 / 34.2	6.24 / 32.8
Proposed 4	Stage II	6.02 / 30.0	6.43 / 35.9	6.25 / 34.0	6.19 / 32.0
Baseline 1	—	7.42 / 48.6	9.29 / 74.0	9.01 / 73.0	8.07 / 55.9
Baseline 2	—	6.45 / 34.7	8.96 / 92.3	8.31 / 87.3	7.82 / 57.2

fine-tuning. Thus, for conciseness, the first-stage results of Proposed 4 are omitted from these tables.

First, let us examine the results for Patient 1. Significant advancements in the proposed systems over the baselines across all the interfered conditions are readily apparent in Tables 4.2, 4.3, and 4.4. In addition to this, when looking at Table 4.2, an interesting finding is noted under the metrics for converting clean EL speech. Here,

owing to the use of low-quality parallel SD for training as well [45], Baseline 2 outperformed Proposed 1, with 6.15 dB versus (vs.) 6.44 dB in MCD, and 32.1% vs. 39.2% in CER. However, Proposed 2, 3, and 4 excelled over Baseline 2, despite leveraging the same volumes of speech data. Moreover, there was a consistent improvement observed from Proposed 1 to 4. It is expected that incorporating a broader range of interferences into the EL SD should assist the model in learning more robust and discriminative features. These are beneficial for identifying linguistic information and thus enhancing the conversion of clean EL inputs. Conversely, Baseline 2 performed worse than Baseline 1 under the interfered input conditions. Using low-quality parallel SD reduces the generalizability of the system to real-world scenarios if there are significantly large environmental mismatches.

Next, the conversion results of tests under interfered EL input conditions are shown in Tables 4.2, 4.3, and 4.4. On this broader evaluation suite, the benefits of the proposed method are clearer, as consistent improvements were observed from Proposed 1 to 4. Among the models shown in the tables, Proposed 4 showed the highest performance in 17 out of the total 22 results (covering MCD and CER metrics) across all eleven conditions, whereas Proposed 3 mainly took the second place. This reinforces the effectiveness of using diverse types of processed EL SD during two-stage fine-tuning to transfer the adaptation knowledge of real-world scenarios. Furthermore, among Proposed 2, 3, and 4, the second-stage fine-tuning mainly optimizes the results of the first stage, where MCD showed a clearer optimization than CER (e.g., for "NR-EL" in Table 4.2, MCD / CER decrease from 6.33 dB / 38.4% to 6.09 dB / 36.7% for Proposed 2, and from 6.30 dB / 37.4% to 6.06 dB / 35.9% and 6.02 dB / 34.6% for Proposed 3 and 4, respectively). Note that although the first-stage fine-tuning promotes recognizing and eliminating non-speech information for the converted speech, it still needs to contend with the misinformation contained in large-scale, low-quality SD, which would compromise the speech quality. Then, the second-stage fine-tuning makes the negative impact of SD negligible, consequently maximizing the advantages of the first-stage fine-tuning.

Next, let us evaluate the overall results of Proposed 2, 3, and 4 in Tables 4.2, 4.3, and 4.4. In general, Proposed 3 outperformed Proposed 2 more often than not for each fine-tuning stage, whereas Proposed 4 advanced further than Proposed 3. From the results in "N-EL" and "R-EL" categories in Table 4.2, the performances of these three revealed consistent improvements compared with those under the noisy-reverberant condition, and nearly matched the performance under the clean EL condition. Tables 4.3 and 4.4 showed a similar trend. Although Proposed 2, 3, and 4 showed degradations in performance when converting noisy-reverberant EL data with stronger noise or reverberation, i.e., at an out-of-range SNR at -5 dB, or a T60 of 1.0 second, they were still significantly better than Proposed 1 and the baselines. Moreover, the conversion performances gradually improved and reached their best as the noise/reverberation intensity decreased, as indicated by the results at a higher SNR or lower T60 in Tables 4.3 and 4.4. The above findings indicate that the proposed methods performed reasonably well under a wide range of interfered conditions and adapted particularly well to a single-interference condition or a noisy-reverberant condition with relatively mild noise/reverberation effects. On the other hand, the overall results of Proposed 2, 3, and 4 were mainly better in the "R-NL" category than in the "N-EL" category in Table 4.2, suggesting that noise more detrimentally impacts EL2SP performance than reverberation. As the reverberation tends to stretch the length of EL speech, the proposed systems, thanks to leveraging the seq2seq framework, are more specialized in handling the issue, whereas noise complicates the speech mapping more directly. Taken together, employing interfered data, which encompasses a broad range of T60s and SNRs, coupled with the effective many-to-one training techniques, enables the proposed methods to recognize varying intensities of reverberation and noise, thereby enhancing their robustness in real-world scenarios.

Furthermore, let us examine the results for Patient 2 in Table 4.5. Baseline 2 exhibited better MCDs for interfered conditions than Baseline 1, which differs from the observations on Patient 1. This can be attributed to the larger size of the original

Table 4.6: Comparison results of baseline systems using SE training methods when the EL input from the Patient 1 dataset is Noisy-Reverberant (NR). Baseline 1 represents the lower bound.

Baseline systems	**NR-EL**
	MCD [dB] / CER [%]
Baseline 1	8.81 / 77.7
E-pt-Base1	9.32 / 74.1
E-ft-Base1	7.95 / 58.7
E-ft-c-Base1	7.70 / 58.6
E-ft-j-Base1	7.61 / 58.3
E-ft-c-d-Base1	7.00 / 50.1
E-ft-j-d-Base1	6.91 / 49.5
E-ft-j-d-Base2	6.34 / 38.8
E-2ft-j-d-Base2	6.25 / 36.9

EL data in Patient 2, which results in higher-quality interfered EL SD, and consequently, improved converted speech quality. However, the higher CERs of Baseline 2 still indicate that imperfections in SD introduce environmental mismatches, negatively impacting real-world conversion intelligibility. More crucially, the findings for the proposed systems align closely with those from Patient 1 (Table 4.2): (1) Across all test conditions, the proposed systems outperformed the baseline systems and handled interfered conditions well, (2) Incorporating SD with two-stage fine-tuning lead to continuous improvements in system performance, ultimately yielding the best-performing model, and (3) Comparing different proposed systems, during the two-stage fine-tuning, introducing augmented data with more fine-grained interference properties further enhanced the model performance, making Proposed 4 the optimal system. By and large, these findings further verify the generalizability and robustness of the proposed methodology across different EL2SP scenarios.

Comparison with seq2seq baselines using extra SE modules: Given that the noisy-reverberant condition represents the most severe challenge, the Patient 1 dataset was leveraged to provide a detailed summary and analysis of the performance differ-

Table 4.7: Evaluation results of the modules based on different SE methods, according to the comparison between processed and Clean EL data from the Patient 1 dataset. The final processed data from Noisy-Reverberant (NR) input after both denoising (dn) and dereverberation (dr) is evaluated. Specifically, the intermediate denoised data, processed by E-pt, E-ft, and E-ft-c is also evaluated. The comparison result between NR inputs and the corresponding clean speech is used as the lower bound.

SE modules	**Comparison**	**Evaluation metrics**	
		SI-SDR [dB]	STOI
—	NR vs. Clean	2.08	0.62
E-pt	dn-NR vs. Clean	2.12	0.59
	dn-dr-NR vs. Clean	1.45	0.57
E-ft	dn-NR vs. Clean	4.42	0.69
	dn-dr-NR vs. Clean	6.65	0.70
E-ft-c	dn-NR vs. Clean	4.42	0.70
	dn-dr-NR vs. Clean	10.09	0.72
E-ft-j	dn-dr-NR vs. Clean	10.25	0.73
E-2ft-j	dn-dr-NR vs. Clean	10.86	0.73

ences among Baseline 1 and the baseline systems using SE modules under this condition, as shown in Table 4.6. Also, as shown in Table 4.7, experiments similar to those outlined by Choi *et al.* [70] were carried out, to thoroughly show the performances of all SE methods used by quantitatively comparing the processed EL data with the initial noisy-reverberant EL inputs in terms of SI-SDR and STOI metrics.

Note that, compared with the lower bound, E-pt presented more negative results in Table 4.7. The quality of the processed speech continued to deteriorate after undergoing both denoising and dereverberation processes. This indicates that the two SE models in E-pt based on normal training data are not generalizable to noisy-reverberant EL inputs. This inadaptability inevitably leads to accumulated errors during enhancement processing. Conversely, E-ft and E-ft-c, both fine-tuned using additionally interfered EL SD, demonstrated improved performance compared with the lower bound, suggesting that even using imperfect SD also aids SE training. E-ft-j exhibited a further enhanced performance for both SI-SDR and STOI, which demonstrated the effective-

ness of the joint training proposed. Furthermore, E-2ft-j, which extends E-ft-j through two-stage fine-tuning, achieved the best performance among all SE methods, with SI-SDR and STOI scores of 10.86 dB and 0.73, respectively, further verifying that the two-stage fine-tuning strategy is effective not only for VC but also for SE training.

In Table 4.6, the poor SE effect of E-pt also affected the conversion results of E-pt-Base1, rendering it even worse than those of Baseline 1. In addition, all systems based on Baseline 1, namely, E-pt-Base1, E-ft-Base1, E-ft-c-Base1, and E-ft-j-Base1, exhibited a progressive optimization trend consistent with the performance of the SE methods documented in Table 4.7, reflecting their intrinsic reliance on the performance of SE modules. However, the improvement of these systems remained limited owing to the fact that Baseline 1 cannot adapt to the processed inputs. In contrast, through the training with processed data, E-ft-c-d-Base1 and E-ft-j-d-Base1 handle this issue effectively. When compared with the proposed systems (see "NR-EL" results in Table 4.2), both unsurprisingly achieved better performance than Proposed 1 in terms of MCD and CER. Surprisingly, although E-ft-j-d-Base2 was further improved by additionally using processed EL SD with the same speech volume as Proposed 2, 3, and 4, it reached the equivalent performance as the first-stage fine-tuning of Proposed 2 (6.34 dB vs. 6.33 dB in MCD, and 38.8% vs. 38.4% in CER), and still clearly underperformed the second-stage fine-tuning of Proposed 2, 3, and 4. Furthermore, thanks to the state-of-the-art SE framework, E-2ft-j-d-Base2 further enhanced performance, achieving MCD and CER of 6.25 dB and 36.9, respectively. However, it still fell short compared to the second-stage fine-tuning of the Proposed 2, 3, and 4 systems.

To further validate the effectiveness of joint-framework-based SE modules, the Patient 2 dataset was also used to develop E-ft-j and E-2ft-j. Similarly, the baselines extended with these SE modules, namely, E-ft-j-d-Base2 and E-2ft-j-d-Base2, were compared. The experimental results are documented in Tables 4.8 and 4.9. The findings are aligned with those of the Patient 1 experiments, with Table 4.8 corresponding to Table 4.7 and Table 4.9 corresponding to Table 4.6, reinforcing the effectiveness of the

Table 4.8: Evaluation results of E-ft-j and E-2ft-j, according to the comparison between processed and clean EL data from the Patient 2 dataset. The result between Noisy-Reverberant (NR) inputs and the corresponding Clean speech is used as the lower bound.

SE modules	Comparison	Evaluation metrics	
		SI-SDR [dB]	STOI
	NR vs. Clean	−0.32	0.69
E-ft-j	dn-dr-NR vs. Clean	10.01	0.86
E-2ft-j	dn-dr-NR vs. Clean	10.10	0.87

Table 4.9: Comparison results of baseline systems using SE training methods when the EL input from the Patient 2 dataset is Noisy-Reverberant (NR). Baseline 1 represents the lower bound.

Baseline systems	NR-EL
	MCD [dB] / CER [%]
Baseline 1	9.29 / 74.0
E-ft-j-d-Base2	6.61 / 41.4
E-2ft-j-d-Base2	6.55 / 39.2

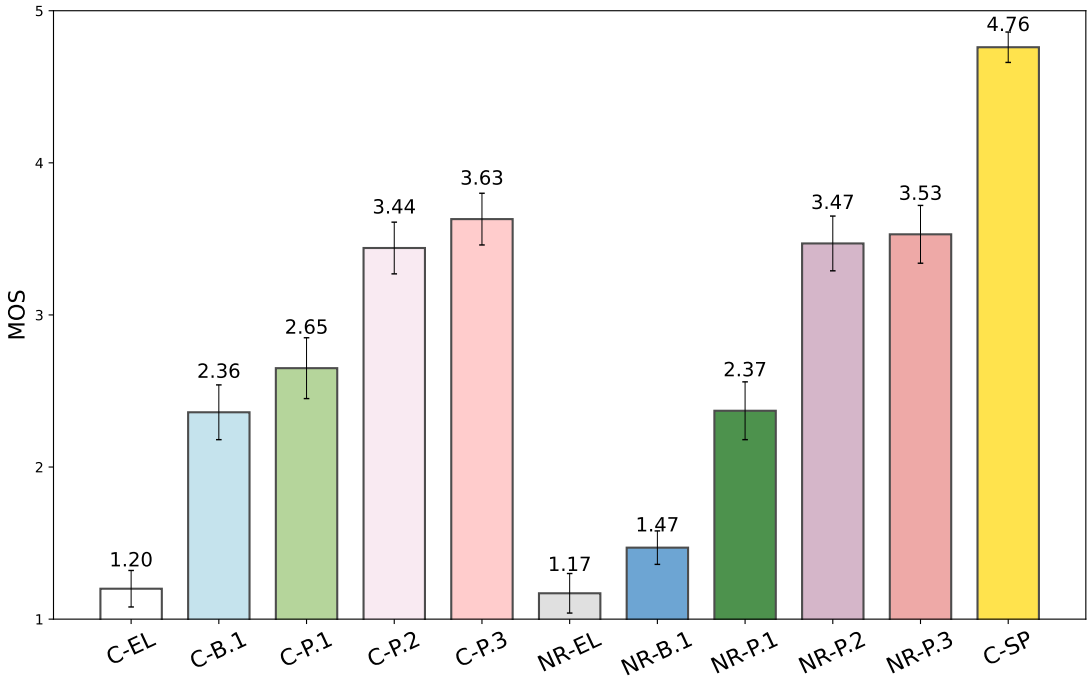
joint framework and the two-stage fine-tuning approach.

The overall comparative study makes the validity of the proposed systems clearer. The seq2seq VC framework is enhanced with more fine-grained interfered SD that represents complex real-world scenarios, leveraging knowledge transfer and error calibration achieved through two-stage fine-tuning. Consequently, the proposed method efficiently achieved the best performance, surpassing the widely used frameworks that rely on extra SE modules.

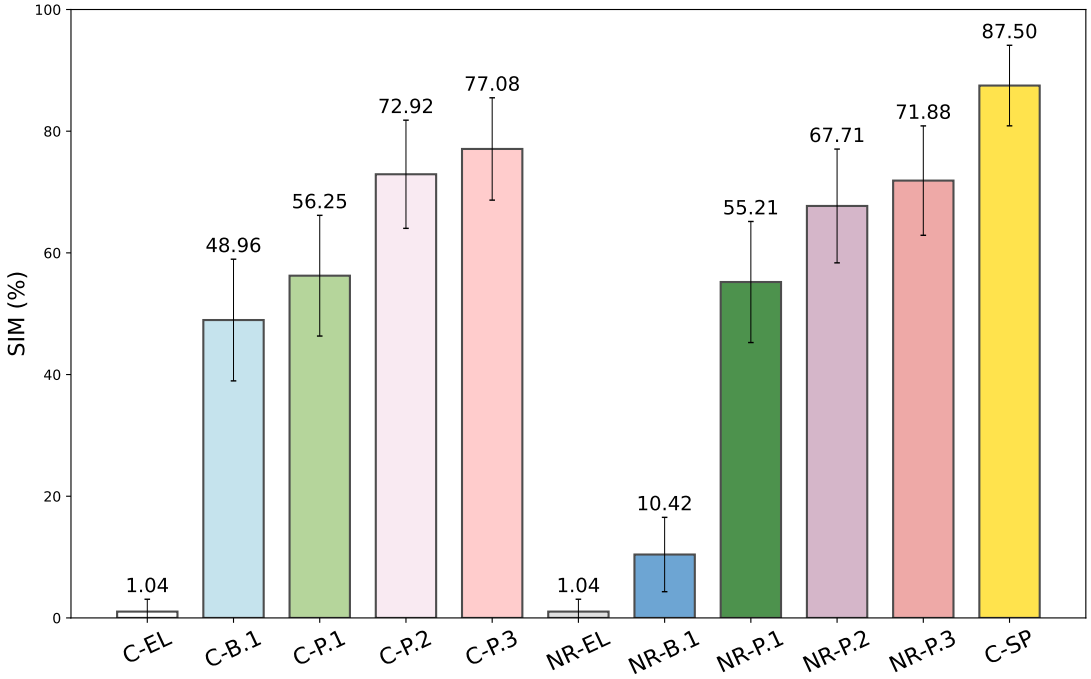
4.4.2 Subjective Evaluation Results

Figure 4.6 shows the MOS and SIM results, incorporating eleven types of speech mixed into the respective test set based on the Patient 1 dataset. Both metrics exhibited a smooth optimization trend in the tasks of converting clean and noisy-reverberant EL speech.

Consistent with the results of the objective evaluations, the proposed systems sig-



(a) MOS results.



(b) SIM results.

Figure 4.6: Mean Opinion Score (MOS) and speaker SIMilarity (SIM) results with 95% confidence interval under Clean (C) and Noisy-Reverberant (NR) conditions, where B.1, P.1, P.2, and P.3 denote the outputs of Baseline 1, Proposed 1, and the second-stage outputs of Proposed 2 and 3, respectively. Clean, Noisy-Reverberant EL, and target normal SPeech are also used as the lower and upper bounds, denoted as C-EL, NR-EL, and C-SP, respectively.

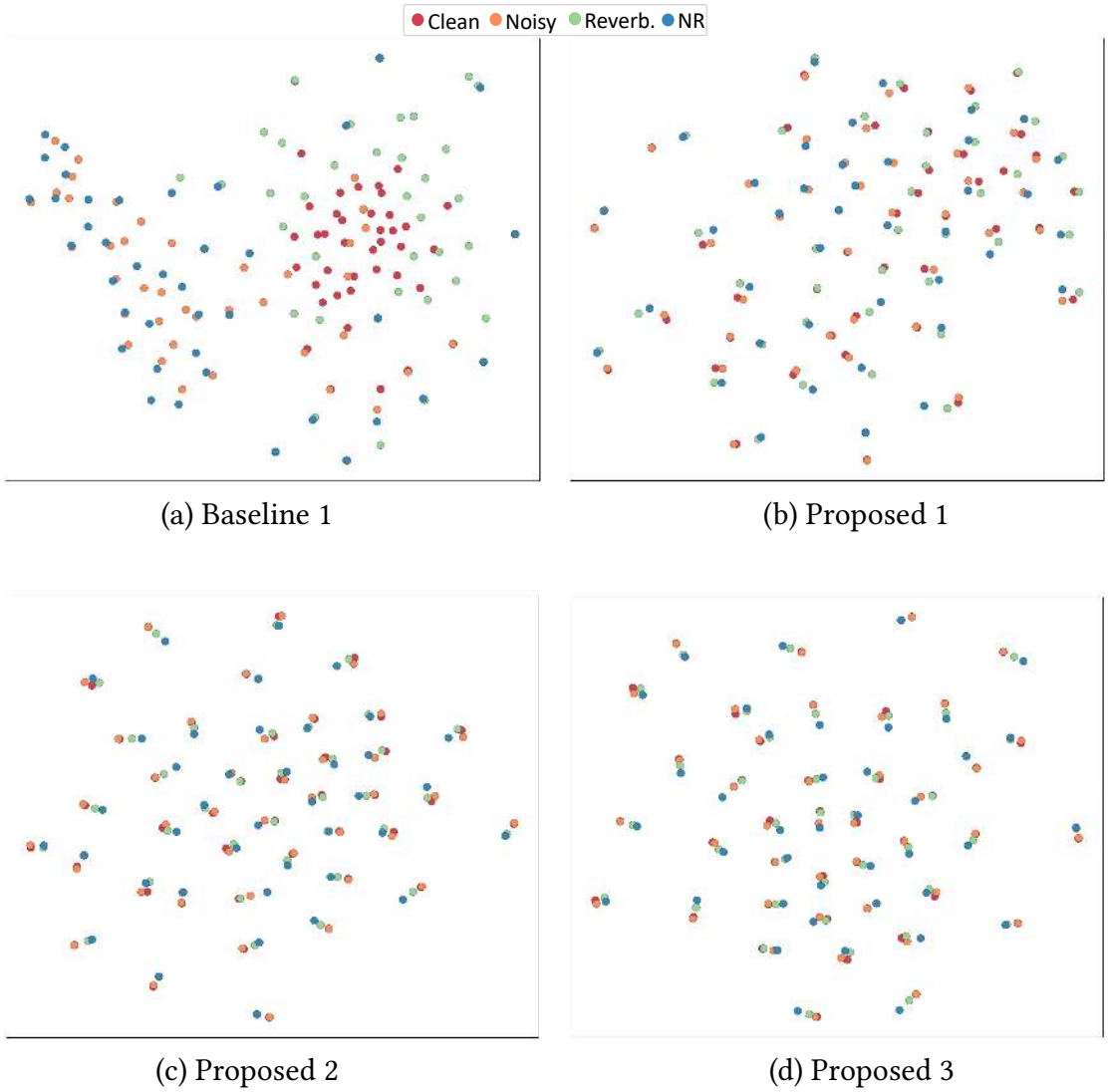
nificantly outperformed Baseline 1, with a progressive enhancement from Proposed 1 to 3. Particularly in noisy-reverberant EL2SP, Proposed 2 and 3 revealed closer speaker similarity and naturalness to the target compared with Proposed 1, underscoring the effective robustness achieved through the two-stage fine-tuning with interfered SD. Moreover, the edge of Proposed 3 over Proposed 2 confirms the benefits of utilizing a broader range of interfered data. On the other hand, the narrow distinctions between these two can be attributed to their utilization of noise and reverberation across varied levels, facilitating robust adaptation to intricate real-world scenarios.

It is surprising to see Proposed 2 and 3 yielded results under the noisy-reverberant condition comparable in naturalness and speaker similarity to those under the clean condition. This showcases the superiority of the proposed systems in converting more natural speech with a closer speaker identity to the target under the noisy-reverberant condition.

4.4.3 Visualizations of the Hidden Representation Spaces

Since the proposed method is expected to assist the model's encoder in filtering out interferences and extracting speech-related knowledge, particularly linguistic representations, visual evidence is provided here by conducting uniform manifold projections [157] to further demonstrate this. Leveraging the Patient 1 dataset, Proposed 1, 2, and 3 were specifically analyzed, and the results of Baseline 1 were used as the lower bound. Hidden representations of the test sets were extracted from the trained encoders of these systems. These representations are then visualized at utterance and phoneme levels, as shown in Figures 4.7 and 4.8, respectively. Note that the encoding results were assessed under four conditions, namely, clean, noisy, reverberant, and noisy-reverberant.

Utterance-level visualization: Besides the impact of environmental interferences on the encoding effects, the difference in linguistic content across utterances is the critical variable, affecting utterance-level representations. In this context, Baseline 1



(a) Baseline 1

(b) Proposed 1

(c) Proposed 2

(d) Proposed 3

Figure 4.7: Visualizations of utterance-level hidden representations extracted from Baseline 1, Proposed 1, and the second-stage fine-tuning of Proposed 2 and 3. Each utterance's frame-wise mean from the latent space is represented as a single dot, and different colors correspond to different conditions. "NR" in labels represents Noisy-Reverberant.

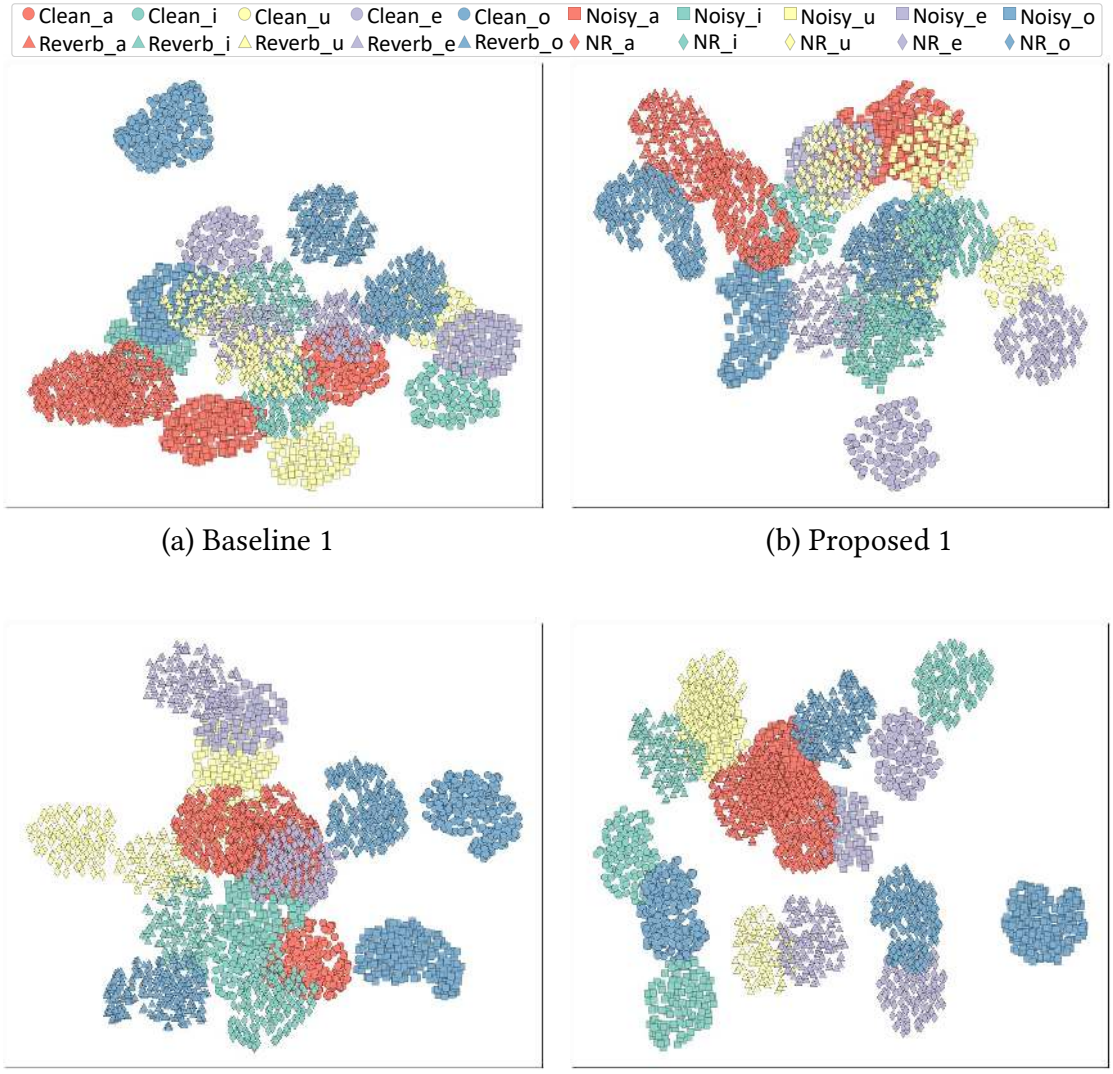
presented a relatively clear clustering effect for clean EL inputs, but the hidden representation space for the interfered EL inputs exhibited poor discriminability. This suggests that Baseline 1 could not adapt to the interfered scenarios.

Proposed 1, 2, and 3 showed roughly similar representation spaces, yet there were notable differences that warrant further analysis. Compared with Baseline 1, Proposed 1 showed better clustering. It is not difficult to infer that, as interferences are eliminated during encoding, speech with the identical linguistic content tended

to cluster closely. Nevertheless, some areas of the hidden representation space still showed weak separation performance, indicating that Proposed 1 struggled to differentiate between features of different sentences owing to its performance limitations and the impact of interferences. Conversely, Proposed 2 and 3 exhibited more effective clustering. Both of them not only achieved more compact clustering for the same utterance contents but also formed well-separated clusters for different utterance contents. These observations closely match the expectation, demonstrating the effectiveness and potential of the proposed method based on the seq2seq architecture.

Phoneme-level visualization: Since this chapter utilizes Japanese datasets, the five most common Japanese vowel phonemes and their corresponding hidden representations are colored to simplify the plots. As shown in Figure 4.8, the phoneme level provides more microscopic and explicit visualization effects than the utterance level. Note that, in each figure, the color of the points indicates the phoneme type, and the shape indicates the environmental condition. Theoretically, an effective representation should clearly cluster the same phonemes regardless of environmental conditions.

For Baseline 1, the five phoneme representations from the clean condition were relatively discretized. However, across all conditions, there was considerable overlap among different representations, suggesting that Baseline 1 does not distinctly differentiate phoneme features. Proposed 1 showed a clearer distribution of the same phonemes, yet overlaps still persisted, albeit slightly improved from Baseline 1. Proposed 2 further exhibited a stronger degree of clustering effect, although the minor overlap reflects the difficulties in distinguishing phonemes with similar pronunciation mechanisms due to interferences (e.g., the phonemes “a” in reverberant speech and “e” in noisy-reverberant speech). Looking at Proposed 3, the visualization of some phoneme representations, such as the phonemes “a” and “u”, showed a high degree of cluster purity, indicating the effective filtration of interferences and enhanced phoneme recognition. In addition, Proposed 3 ensures that the same phoneme types under different conditions cluster closely, and the overlaps between various phoneme



(a) Baseline 1

(b) Proposed 1

(c) Proposed 2

(d) Proposed 3

Figure 4.8: Visualizations of phoneme-level hidden representations extracted from Baseline 1, Proposed 1, and the second-stage fine-tuning of Proposed 2 and 3. The representations of the five Japanese vowels, "/a/", "/i/", "/u/", "/e/", and "/o/", are plotted. In total, 20 types of phoneme representations are visualized (5 phonemes × 4 environmental conditions), distinguished by the colors and shapes of the dots. "NR" in labels represents noisy-reverberant.

representations are almost negligible. It demonstrated the clear delineation and robust grouping for the phonetic features of EL inputs across environmental variations. However, note that Proposed 2 and 3 do not consistently conform to the above theoretical assumption, in that a number of the same phoneme representations do not form compact clusters owing to the different acoustic properties between interfered and clean EL speech. This also reflects the performance difference of the proposed

systems when converting interfered input compared with clean input.

Overall, although capturing subtleties between phonemes poses a greater challenge than utterance-level representations, the notable improvement from Proposed 1 to 3 corroborates their capability to identify and filter interferences, underscoring the overall efficacy of the proposed method.

4.5 Conclusions

This chapter developed and evaluated a training strategy within a seq2seq framework to simultaneously address (1) limited data and (2) lack of robustness against real-world interferences in EL2SP. Building on the encoding mechanism of seq2seq VC and the insights from transfer learning, the paradigm introduced in Chapter 3 was extended, particularly its fine-tuning techniques, by developing a unified two-stage fine-tuning framework tailored for realistic conditions. This framework leverages readily available, low-quality parallel SD, proven effective in low-resource scenarios, and further augments it with diverse and fine-grained interfered conditions to construct many-to-one mappings toward clean normal speech. Based on this framework, the model's generalization was improved to real-world scenarios. Moreover, the two-stage fine-tuning not only inherits beneficial information but also diminishes the negative impact of SD, thus achieving optimal performance.

On the basis of the flexibility of the proposed training framework, multiple systems were designed and evaluated at different fine-tuning levels. Experimental results demonstrated that the proposed systems outperformed the baseline systems trained on clean data in the EL2SP task under different test conditions. Furthermore, the proposed systems were systematically compared with the mainstream approach using external SE modules. Although the performance of these baseline systems can be continuously improved by optimizing the SE architectures, the best-performing system in the proposed method still prevailed in direct comparisons.

In addition to outperforming the baselines, another advantage of the proposed method manifests in its simpler architecture. Adding new modules or architectures often increases complexity and separately handles interference elimination and conversion, while also depending on considerable module-specific training data. The above factors pose difficulties for actualization in practical scenarios. In contrast, the proposed systems rely solely on a single seq2seq architecture. Moreover, the proposed fine-tuning method using SD is efficient and easily applicable in real-world environments. Encoder analyses further revealed that the proposed systems are capable of disentangling linguistic information from interfered signals.

Taken together, this chapter demonstrated that the proposed training method provides a practical and effective solution for interference-robust EL2SP under low-resource settings. It highlighted the potential of seq2seq architecture, when properly fine-tuned with synthetically interfered data, to achieve both high-quality and interference-resilient EL2SP. Therefore, this work provided a relatively fresh perspective for readers, promoting cost-effective training approach based on advanced model architectures that enhance adaptation to downstream tasks with limited resources, and the capability to disentangle non-essential knowledge. On the other hand, the potential of seq2seq VC in processing multi-source inputs naturally makes us consider that additional types of useful information may benefit representation learning, which will be further explored in the next chapter.

5 | Representation Learning Method Integrating Text and Speech Representations

This chapter proposes a novel representation learning method that incorporates auxiliary text information into a single sequence-to-sequence (seq2seq) Voice Conversion (VC) framework to improve the performance of ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP). Up to this point, the optimization strategies for EL2SP have been largely limited to mappings between speech features. While some representative transfer learning methods have achieved steady improvements over the conventional frame-wise approach, the substantial acoustic mismatch between EL and normal speech still introduces inevitable inaccuracies, leaving performance bottlenecks unresolved, as discussed in Subsection 1.3.3.

To address this issue, the proposed method explicitly integrates paired EL speech and text representations, enabling more effective mapping to target speech. Specifically, text information is incorporated to enrich intermediate representations and guide alignment learning between the speech encoder and decoder. Next, an autoencoder-style reconstruction training (updating only the encoder) is further introduced to allow the final EL2SP model to inherit the integrated representations without adding system complexity. Experimental results demonstrate that the proposed method, when combined with data augmentation, clearly outperforms baselines that rely solely on speech

representations. Notably, the proposed method achieves a 19.2% reduction in Character Error Rate (CER), and a 0.67 increase in Mean Opinion Score (MOS) of naturalness compared with a typical pretraining–fine-tuning baseline, confirming its effectiveness.

The organization of this chapter is as follows. Section 5.1 provides an overview of this chapter. Section 5.2 presents the proposed speech–text representation learning method, consisting of three parts. Section 5.3 describes the experimental evaluation settings, followed by the experimental results in Section 5.4. Finally, Section 5.5 concludes this chapter.

5.1 Introduction

This section summarizes the motivation and contributions of the proposed method. As discussed earlier, current seq2seq methods for EL2SP mainly rely on speech-only transfer learning strategies to address the low-resource challenges. Representative approaches include the typical pretraining–fine-tuning approach [40], which first pretrains a model on a large-scale normal corpus and then fine-tunes it on a small-scale EL2SP dataset, as well as Text-To-Speech (TTS)-oriented pretraining followed by two-stage fine-tuning incorporating low-quality Synthetic Data (SD). Both strategies achieve steady improvements over conventional frame-wise VC, as demonstrated in Chapter 3. Nevertheless, their performance remains limited due to the substantial acoustic mismatches between EL and normal speech.

Meanwhile, findings from Chapters 3 and 4 further indicate two insights: (1) Potential of seq2seq modeling to handle multi-source input tasks, and (2) Crucial role of stable, linguistically enriched intermediate representations. As emphasized in Subsection 1.3.4, compared with speech, text information is inherently cleaner and offers more direct cues for alignment in representation learning. Related studies reviewed in Subsection 2.1.6 also support this notion, showing that incorporating the power of TTS models provides potential directions to more easily extract pure linguistic inter-

mediate representations, thereby promoting more accurate encodings. For example, utilizing TTS-related modules or priors has been shown to improve modeling quality in VC by Zhang *et al.* [79] and Park *et al.* [81]. However, these studies have fundamental limitations. First, they increase the complexity of the original VC framework, as the former imposes high demands on the model size and hyperparameter design to support a generalized attention adaptable to different tasks, while the latter relies on text as input during inference. These factors hinder their practical applicability. Furthermore, as emphasized in Section 2.3, text has typically been used indirectly as *a priori* information or alignment reference, leaving its potential open for further investigation. Finally, these methods have not been validated in the EL2SP task with a more challenging mapping condition.

Motivated by these observations and inspired by the fusion strategies discussed in Section 2.3, this chapter proposes a seq2seq-based representation learning method. Unlike speech-feature-based representations alone, the proposed method explicitly integrates EL speech with its text encodings to enhance EL2SP within a single VC framework. First, a unified training network is designed, comprising a speech encoder, a text encoder, and a speech decoder. During training, the network learns a richer and more precise intermediate representation space by jointly leveraging speech and text inputs. The proposed method then effectively inherits these representations to improve EL2SP, while ultimately preserving the simplicity of the original seq2seq VC framework.

The key contributions of this chapter are summarized as follows:

- This is the first work to apply a joint network integrating text and speech features to EL2SP. A text encoder and a decoder, derived from a TTS model fine-tuned on target normal speech, are introduced. It can generate pure-linguistic representations for target speech mapping and thus facilitates smoother EL-to-normal speech mapping.
- A straightforward yet effective autoencoder-style training strategy is proposed.

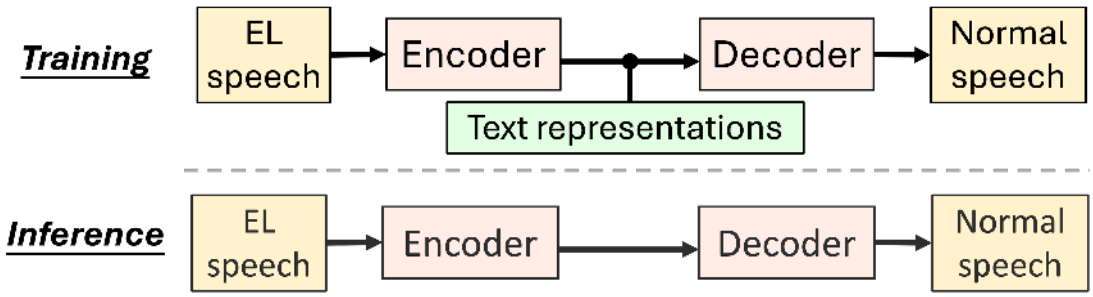


Figure 5.1: Basic concept of the proposed representation learning method.

After removing the text encoder, the speech encoder can approximate the joint representation space achieved by combining text and speech features, finalizing enhanced EL2SP within a single VC framework.

- By leveraging data augmentation with SD and the corresponding text data, the proposed system significantly outperforms prior works [40], [45] across both objective and subjective evaluations.

5.2 Speech–text Representation Learning Method

Before detailing the proposed method, a conceptual illustration is provided here to help readers better grasp the core idea. As shown in Figure 5.1, the proposed method introduces text representations into the training process, enabling the model to learn and retain enriched joint representations while preserving the simplicity of the seq2seq architecture and its inference procedure.

In particular, as illustrated in Figure 5.2, the proposed method comprises three parts: *Part 1* involves preparing pretrained modules using the TTS and EL2SP datasets for initializing the joint network (Subsection 5.2.1). *Part 2* focuses on joint network training to learn intermediate representations containing speech and text information from the EL2SP dataset (Subsection 5.2.2). *Part 3* entails reconstruction training using only speech data of the EL2SP dataset to finalize the EL2SP model (Subsection 5.2.3).

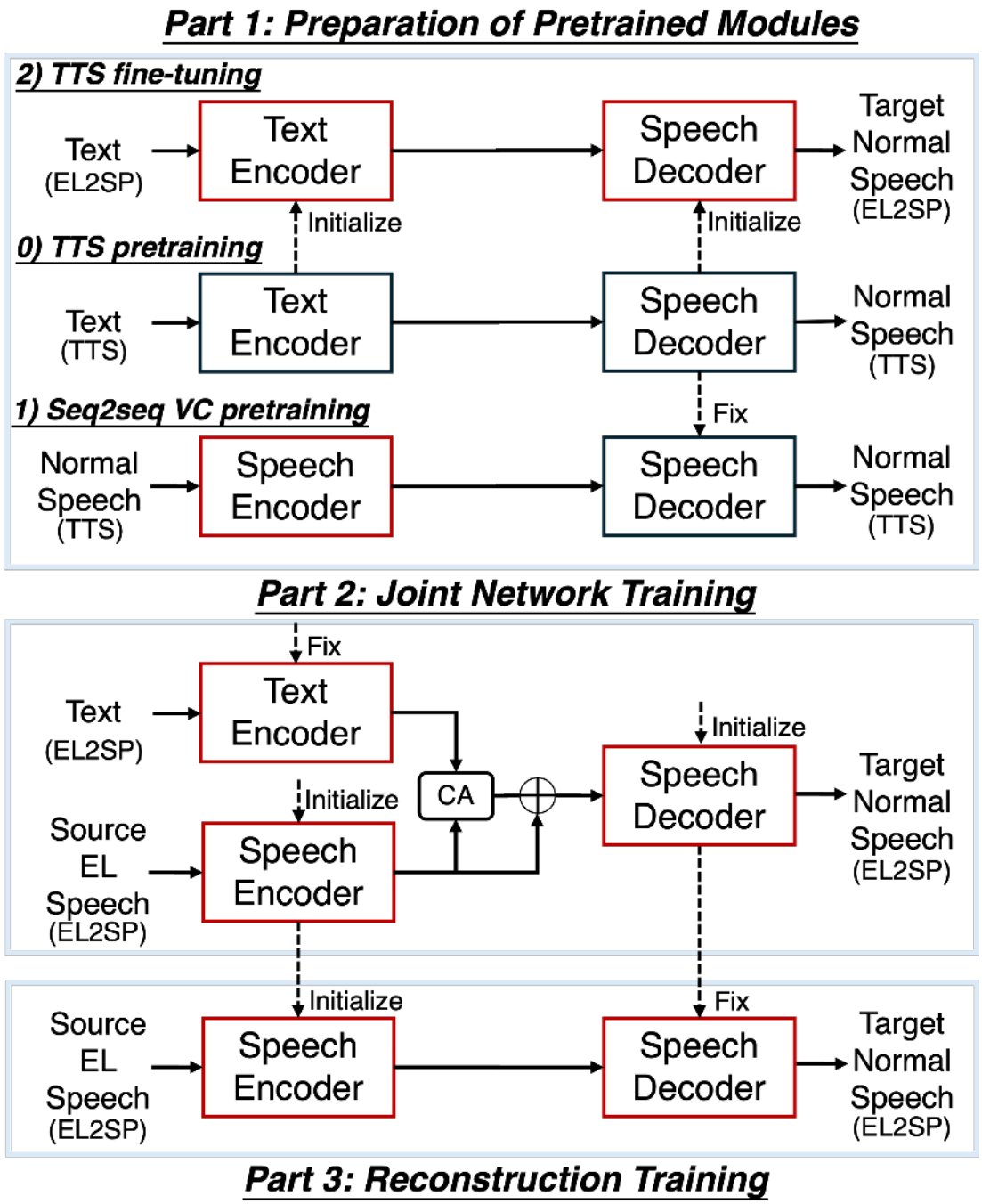


Figure 5.2: Overview of the training framework for the proposed representation learning method. Modules trained in *Part 1*, namely the text encoder, speech encoder, and speech decoder (highlighted in red blocks), are inherited by the joint network in *Part 2*. CA denotes the Cross-Attention block. In each training step, the corresponding training dataset is marked in brackets.

5.2.1 Part 1: Preparation of Pretrained Modules

(1) Pretrained seq2seq VC model, and (2) phoneme-level TTS model for target normal speech of EL2SP are constructed to prepare pretrained modules for *Part 2*. Both are initialized using the Transformer-based [87] TTS pretraining from a normal-speaker TTS database and follow the training procedure described in Chapter 3. Therefore, the detailed description for these two steps are omitted here.

Note that in Step 2) of *Part 1*, the pretrained TTS model is fine-tuned using the target normal speech data and the corresponding transcriptions from the EL2SP dataset. While the limited data restrict the TTS performance, it remains capable of essential mapping from text representations to the target normal speech.

5.2.2 Part 2: Joint Network Training for Integrating Text- and Speech-based Representations

The parameters of the text encoder and speech decoder from the pretrained TTS model, along with the speech encoder from the pretrained VC model, are utilized to initialize the new joint network. Here, EL speech data and text data serve as dual inputs, while the target normal speech data are used as output. Since the text encoder processes the same text features as in *Part 1*, it is frozen during training. Instead, training focuses on updating the speech encoder and decoder to adapt to EL input and to the fused representations of text and speech, respectively.

- **Text Representations.** Let $T = (t_1, t_2, ..., t_l)$ represent a text feature sequence, where l denotes its length. The sequence is fed into the text encoder and mapped as text representations H_T, which can be simplified as:

$$H_T = \text{TextEnc}(T) \in \mathbb{R}^{l \times d}, \tag{5.1}$$

  where d denotes the dimension of the representations.

- **Speech Representations.** Given a speech sequence $S = (s_1, s_2, ..., s_n)$, where n denotes its length in frames, the speech representations H_s are processed by the speech encoder as:

$$H_S = \text{SpeechEnc}(S) \in \mathbb{R}^{n \times d}. \tag{5.2}$$

- **Integrated Representations.** After generating H_T and H_S, they are combined using a cross-attention block, formulated in the same manner as the Multi-Head Attention (MHA) mechanism described in Eqs. (2.5)–(2.7). In this setting, H_S acts as the query, while H_T from the frozen text encoder serves as both the key and value. For each acoustic time step, the cross-attention module computes a weighted combination of text sequence based on attention scores, resulting in a soft semantic alignment between speech and text representations. Subsequently, a residual connection is applied to reintegrate H_S considering it as a speech-dominated mapping process. Hence, the overall integrated representations H_C capture both acoustic details and aligned textual semantics, which can be expressed as:

$$H_C = \text{MHA}(H_S, H_T, H_T) + H_S \in \mathbb{R}^{n \times d}. \tag{5.3}$$

This process ensures efficient training, during which the construction of H_C is guided by contextual information from H_T while preserving the complete H_S. In addition, the training objective defined using the target speech features can somewhat guide the cross-attention mechanism during training. Thus, the model is encouraged to attend to and integrate those textual semantic components that are beneficial for the speech decoding task. Accordingly, the updated speech decoder has adapted to H_C and thus can decode it into the target speech.

5.2.3 Part 3: Reconstruction Training for ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP) Framework

Part 3 aims to inherit the integrated representations H_C and adapt the network to the EL2SP scenario without relying on text inputs. To achieve this, the text encoder is removed and the speech encoder and decoder are initialized using the parameters from *Part 2*. The model takes the same speech input S. Inspired by Voice Transformer Network (VTN) [82], an autoencoder-style training approach is employed by fixing the decoder and updating only the speech encoder.

Since the decoder is frozen, the characteristics of the intermediate representations it accepts remain unchanged. As previously described, the decoder has already adapted to H_C during *Part 2*. Therefore, to align with the decoder's mapping requirements, the speech encoder is forced to update to reconstruct representations $\tilde{H}_S$, that closely approximate H_C when only consuming S as:

$$\tilde{H}_S = \text{SpeechEnc}(S) \in \mathbb{R}^{n \times d} \approx H_C \in \mathbb{R}^{n \times d}. \tag{5.4}$$

Here, the original seq2seq training objective l_{total} defined in Eq. (2.14) is extended by introducing an additional alignment term. The final loss function in *Part 3* is then expressed as:

$$l = l_{\text{total}} + \lambda l_1(\tilde{H}_S, H_C), \tag{5.5}$$

where λ is a predefined coefficient controlling the strength of the auxiliary alignment. Let us further consider two options for the final system design based on this unified formulation:

- *Option 1*: $\lambda = 0$, i.e., equivalent to the original loss in Eq. (2.14), relying solely on

the reconstruction training process.

- *Option 2*: $\lambda > 0$, i.e., reconstruction training augmented with the integrated representation-guided auxiliary loss by attending to H_C from *Part 2*, for more explicit alignment learning of speech encoder.

Investigation on the impact of *Options 1* and *2* will be fleshed out in the next section.

In summary, this method enables the transfer of more effective integrated representations to the EL2SP framework, thereby enhancing the overall performance without increasing the architectural complexity.

5.3 Experimental Evaluation Settings

5.3.1 Systems

Based on the proposed method in Section 5.2, three systems were designed. To demonstrate their effectiveness, they were compared with the two baseline systems, both focusing on speech-based representations. For a fair comparison, all the systems used the same seq2seq VC pretraining. The respective methodologies and data types used for these comparable systems are documented in Table 5.1.

- *Baseline 1:* It employs a typical pretraining–fine-tuning process as in Yen *et al.*'s work [40], where the pretrained seq2seq VC model is directly fine-tuned using the original EL2SP dataset.

- *Baseline 2:* It follows the methodology introduced in Chapter 3, incorporating the two-stage fine-tuning method using parallel SD for data augmentation. To obtain parallel SD, aside from the target normal TTS model (consistent with *Part 1* in Subsection 5.2.1), the same pretrained TTS model is fine-tuned on the original EL speech data, hence building the EL TTS model. Then, an external text dataset is input into both TTS models to produce larger-scale parallel SD.

Table 5.1: Comparable systems under speech- and integrated-representation-based categories, with the former containing Baseline 1 and 2 systems and the latter comprising Proposed 1, 2, and 3 systems, each equipped with specific training data and methods.

Comparative categories				
Speech-representation-based		**Integrated-representation-based**		
Baseline 1 [40]	Baseline 2	Proposed 1	Proposed 2	Proposed 3
Pretraining & fine-tuning	Pretraining & two-stage fine-tuning	Parts 1 to 3	Parts 1 to 3	Parts 1 to 3 & extended losses
(1) Original data	(1) Original data (2) Synthetic data	(1) Original data (3) Text data	(1) Original data (2) Synthetic data (3) Text data	(1) Original data (2) Synthetic data (3) Text data

During the first-stage fine-tuning, the original EL2SP data and parallel SD are pooled together for training. As noted in Chapter 3, since the SD is in low-quality, the second-stage fine-tuning uses only the original EL2SP data to refine parameters and achieve optimal performance.

- *Proposed 1:* It fully implements *Parts 1* to *3* in Section 5.2, which serves as a standard version of the proposed method. As shown in Table 5.1, only the original EL2SP dataset with its paired text data are used to conduct *Parts 2* and *3*. Meanwhile, *Option 1* ($\lambda = 0$) is selected as in Subsection 5.2.3, to make *Part 3* retrain the original loss function l_{total}.

- *Proposed 2:* Inspired by the use of imperfect SD in Chapters 3 and 4, SD could somewhat benefit this method as well. Thus, in *Part 2*, both the original EL2SP dataset and parallel SD, along with their corresponding transcriptions, are used to train the joint network. In *Part 3*, since SD still contains considerable inaccuracies that may affect the final performance, only the original EL2SP dataset is used for reconstruction training. Furthermore, the loss function again follows *Option 1*. Therefore, it can be viewed as an further enhanced version of the proposed method.

- *Proposed 3:* It shares the same training process and input data as *Proposed 2*,

including the use of both original EL2SP data and parallel SD with their transcriptions for *Part 2*, and the original speech data along with paired text data for *Part 3*. The difference lies in the loss function during *Part 3*, where *Option 2* is adopted by introducing the integrated representations H_C generated in *Part 2* as an auxiliary supervision signal, weighted by the coefficient $\lambda = 0.01$. This auxiliary loss is expected to encourage the speech encoder to produce representations that align more closely with H_C, thereby injecting beneficial semantic structure into the encoder learning. λ is set to a relatively small value (0.01) to maintain the dominance of the original reconstruction objective, while the auxiliary supervision acts as a light semantic constraint. This is analogous to offering the model optional guidance inspired by strong exemplars, enhancing training effectiveness without introducing text-based dependency during inference.

5.3.2 Datasets

The dataset construction in this experiment largely follows that described in Chapter 3, where the pretrained TTS and seq2seq VC models were built using the Japanese JSUT corpus [141] with 7,296 utterances, totaling around 10 hours. Note that the corresponding text data were also used to generate SD and contribute to the joint network training in Subsection 5.2.2.

To evaluate the proposed method, a small-scale Japanese EL2SP dataset was constructed. A male laryngectomee recorded 200 utterances of EL speech (around 10 minutes), while a healthy speaker recorded 413 utterances of normal speech (around 20 minutes). This EL2SP dataset constituted a semi-parallel corpus, as the EL speech utterances were a subset of the normal speech set.

For system development, 20 parallel utterances were used as the development set and 40 as the test set. Baseline 1 and Proposed 1 utilized only the remaining parallel utterances (140) for training. The first-stage fine-tuning of Baseline 2 and the joint network training of Proposed 2 and 3 incorporated TTS-generated parallel SD

(7,296). Meanwhile, to make full use of the original training corpus (353), they were supplemented with the corresponding EL SD to the extra portion of the original normal speech during these stages. Then, their final stages only used the original parallel data (140) for training.

5.3.3 Implementations

All the systems were implemented with the ESPnet [142] toolkit, where speech signals were consistently resampled at 24 kHz, and the 80-dimensional mel filterbanks with 2,048 Fast Fourier Transformation (FFT) points and a 300-point frame-shift were used to extract acoustic features. The base configurations for the TTS and seq2seq VC models involved were similar to the official hyperparameter settings, using six-layer blocks with four attention heads in the encoders and decoders as mentioned in Subsection 2.1.3, while the reduction factor was set to 3 and the gradient accumulation steps to 1. Additionally, the cross-attention block in the joint network also used four attention heads. The input text feature sequences for the TTS models contained Japanese-based phonemes and pause information. The learning rate was set to 0.001, with the LAMB [143] and Noam [87] optimizers for VC and TTS, respectively. The finalized models from the proposed methods kept the basic VC framework, thus having identical model sizes of 30.4 million trainable parameters as the baselines. To reconstruct waveforms of parallel SD and EL2SP outputs, two speech-dependent Parallel WaveGAN neural vocoders [144] were trained separately for EL and normal speech, following the original configurations.

5.3.4 Evaluation Metrics

5.3.4.1 Objective Evaluation

Mel-Cepstrum Distortion (MCD, in dB), Character Error Rate (CER, in %), log F0 Root Mean Square Error (F0 RMSE), and log F0 CORRelation (F0 CORR) were used to

Table 5.2: Objective evaluation results on the converted speech from the five comparable systems.

Systems	Objective Evaluation Metrics			
	MCD [dB] (↓)	**CER** [%] (↓)	**F0 RMSE** (↓)	**F0 CORR** (↑)
Baseline 1 [40]	7.17	41.3	0.26	0.65
Baseline 2	6.24	23.3	0.25	0.67
Proposed 1	6.21	31.5	0.24	0.67
Proposed 2	6.04	22.1	0.23	0.69
Proposed 3	6.00	20.7	0.22	0.69

evaluate various aspects of the converted speech, as defined in Subsection 2.4.1.

5.3.4.2 Subjective Evaluation

Mean Opinion Score (MOS) test, as defined in Subsection 2.4.2 and accordant to the evaluation setup in Chapters 3 and 4, was conducted, to rate naturalness of the converted samples. Twenty-five native Japanese speakers were recruited. Fifteen samples from each system were randomly chosen for each listener. Audio samples are available online[1].

5.4 Experimental Results

5.4.1 Objective Evaluation Results

Table 5.2 summarizes the overall results. First, let us compare the Baseline 1 and Proposed 1 systems, as they use the training data of the same scale. Proposed 1 significantly outperformed Baseline 1, especially in MCD and CER. This demonstrates that the method of learning integrated representations is more effective than that of learning speech-only representations.

Next, let us compare Proposed 1 with the Baseline 2 system. Owing to the two-stage fine-tuning incorporating much larger-scale parallel SD, Baseline 2 clearly ex-

[1] https://silenticymoon.github.io/EMBC2025demo/

celled over Proposed 1 regarding CER. This indicates that even low-quality SD still retains rich knowledge to improve conversion intelligibility. However, it is notable that despite being trained on the limited original data, Proposed 1 performed comparably to or even slightly better than Baseline 2 in the other three metrics, highlighting the robustness of the proposed method. Moreover, this suggests that although Baseline 2 attempts to correct the potential errors from SD during fine-tuning, its reliance on speech-only representations still introduces trade-offs.

Compared with the Proposed 1 and both baseline systems, the Proposed 2 system achieved consistently better performance across all four metrics. In addition to the richer information gain from the increased data volume, the presence of text representations provides an important guidance during mapping. Specifically, text information helps filter out irrelevant/noisy variations in SD and facilitates the reconstruction of more accurate integrated representations, thereby enhancing the overall performance.

Finally, the Proposed 3 system further improved performance compared with Proposed 2, achieving the best results on the four metrics among all proposed systems. In particular, we can observe a clear reduction in CER, confirming the effectiveness of introducing the additional text-guided loss in *Part 3*. Since this loss serves as a semantic constraint, when approximating to integrated representations during training, it facilitates the speech encoder in capturing higher-quality linguistic information, which is closely related to the conversion intelligibility, from speech-only inputs. Moreover, assigning a relatively small weight ($\lambda = 0.01$) proved sufficient to guide the reconstruction process toward generating the target speech with higher intelligibility. On the other hand, the gains of Proposed 3 in MCD, F0 RMSE, and F0 CORR were marginal. We can attribute this to the very limited amount of speech data available for reconstruction training, which constrains the extent of improvements in spectral quality.

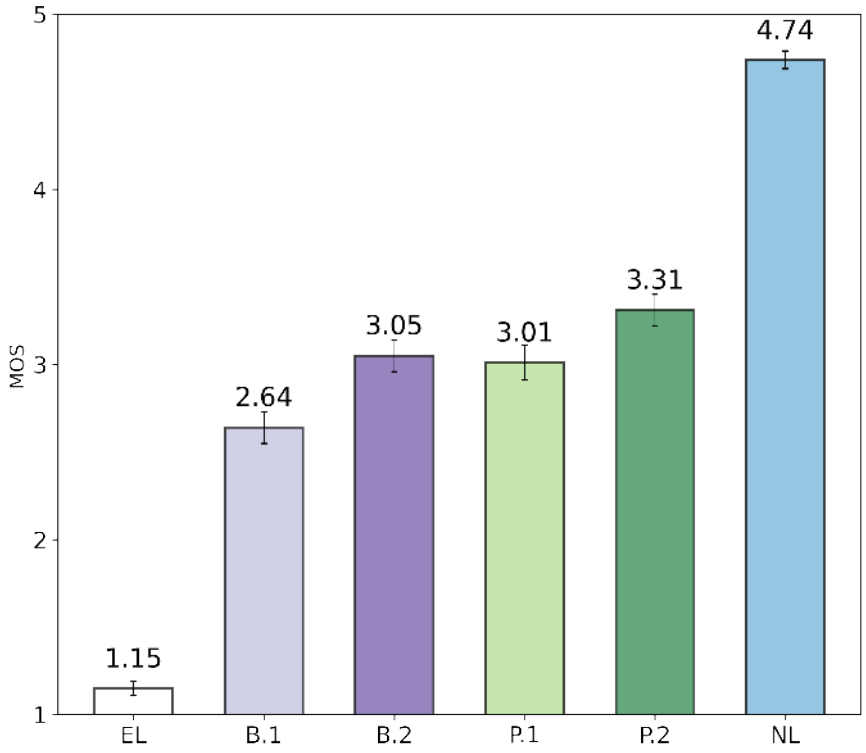


Figure 5.3: Mean Opinion Score (MOS) results with 95% confidence interval, where EL, NL, B.1, B.2, P.1, and P.2 denote EL, normal speech, and the final conversion of Baseline 1, Baseline 2, Proposed 1, and Proposed 2, respectively.

5.4.2 Subjective Evaluation Results

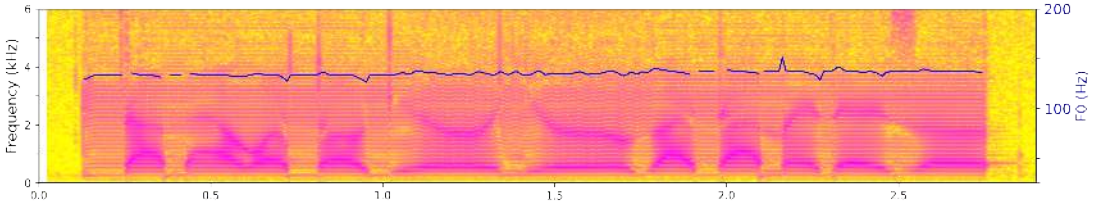
As shown in Figure 5.3, the outcomes of Baselines 1 and 2 and Proposed 1, 2, and 3 were randomly selected, accompanied by their original EL and normal counterparts, for the subjective test. The results indicated that all comparable systems yielded significantly more natural speech than the original EL speech. Among them, Proposed 1 performed notably better than Baseline 1 and reached to a close level to Baseline 2. This suggests that although the outcome of Proposed 1 was less intelligible than that of Baseline 2 (as shown in Table 5.2), it still revealed a relatively high level of naturalness in terms of factors such as fluency and prosody. Moreover, Proposed 2 obviously outperformed both baselines, exhibiting the closest naturalness to the normal speech among all compared systems. These findings align with the trends in the objective results, reinforcing the effectiveness of the proposed framework incorporating text

representations.

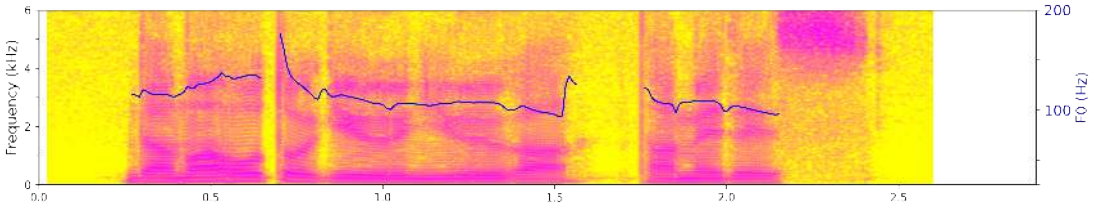
5.4.3 Spectrogram Analysis

Figure 5.4 visualizes the spectrograms and F0 contours of samples from Baseline 2 and Proposed 1, 2, and 3, with EL and normal speech serving as lower and upper bounds, respectively. As expected, compared with normal speech, EL speech exhibited a spectral structure that lacked detail and had minimal pitch variation, as reflected in its flat F0 contour. In addition, EL speech had the longest duration due to its mechanical pronunciation, as noted in Section 1.1. Furthermore, all four systems under comparison clearly enhanced the EL speech. In particular, Baseline 2 and Proposed 1 yielded roughly similar spectral and F0 patterns, while Proposed 1 showed slightly worse F0 contour shape and duration. This observation is consistent with the objective and subjective results. It highlights the robustness of the proposed method with limited data, performing closely to the speech-representation-based approach that uses larger-scale data augmentation.

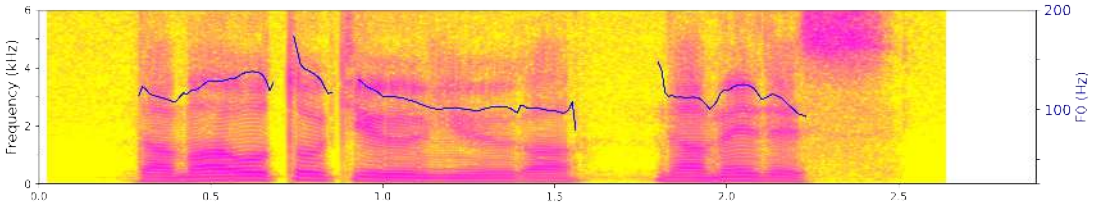
Moreover, both Proposed 2 and 3 further enhanced EL speech, generating spectral structures and F0 trends that more closely resembled those of normal speech. In particular, both systems revealed better pause and duration control compared with Baseline 2 and Proposed 1, resulting in speech rates most closely matching natural pronunciation. These findings suggest that the integration of text representations enriches intermediate features, promoting a accurate reconstruction process. Finally, it is reasonable to see the outputs of Proposed 2 and 3 were highly similar, since this observation is in line with their trivial differences regarding MCD, F0 RMSE, and F0 CORR in the objective metrics.



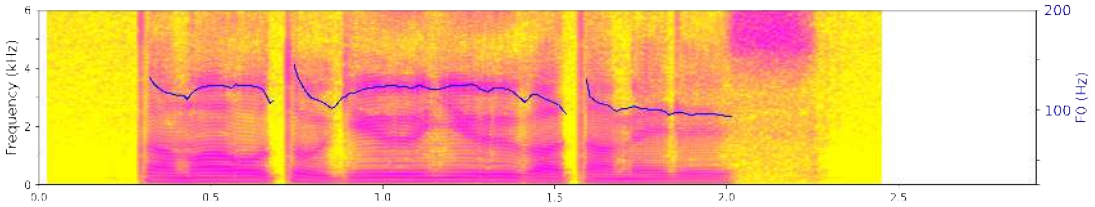
(a) Source EL.



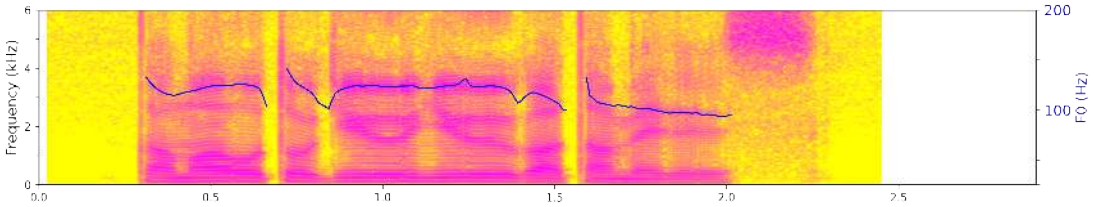
(b) Baseline 2.



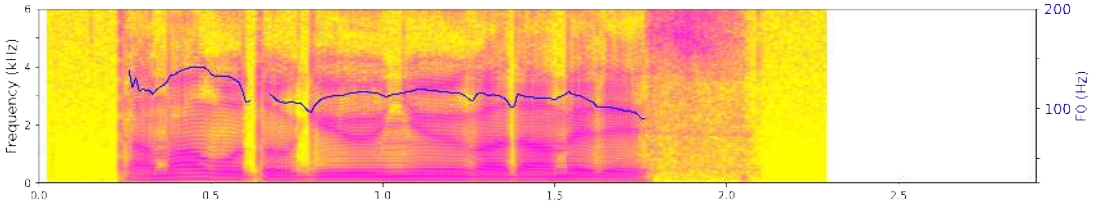
(c) Proposed 1.



(d) Proposed 2.



(e) Proposed 3.



(f) Target normal.

Figure 5.4: Spectrograms and F0 contour of an utterance conversion ‘ta ma go to gyuu nyuu ga da me de su’. The Horizontal axis displays time in seconds. The vertical axis offers the spectral and F0 frequency.

5.5 Conclusions

This Chapter proposed a novel representation learning framework, which integrates text and speech representations within a seq2seq VC framework, to enhance EL2SP performance. Building on the foundations of seq2seq VC and transfer learning, the speech-to-speech approach taken in Chapters 3 and 4 were extended by designing (1) a joint network training and (2) a reconstruction training strategy. Experimental evaluation results clearly indicated that the proposed method consistently outperformed speech-representation-based training strategies. In particular, compared with a typical pretraining–fine-tuning baseline, the proposed method achieved significantly better performance, and it reached a level close to that of a baseline system that used large-scale data augmentation. Furthermore, when further combined with data augmentation, the proposed method achieved optimal performance, winning across all objective and subjective evaluations.

This method offers two key benefits. First, unlike the previous approach that relies on indirect strategies such as using text representations as auxiliary priors or aligning unimodal features, the proposed method explicitly unifies multimodal representations containing text and speech encodings into shared integrated space during joint network training, thereby capturing richer linguistic cues. Second, through reconstruction training, the final system directly forces the encoder to align EL-speech-only representations with these integrated representations, enabling the model to inherit the richer representations for a more accurate mapping between EL and normal speech without increasing network complexity. This is a novel attempt in seq2seq EL2SP modeling, and it reflects that seq2seq models benefit from enriched representations from multiple inputs as well. Moreover, the framework is advanced by refining the reconstruction loss function, reintroducing the integrated representations as an auxiliary term and hence getting a slightly optimized performance, highlighting loss design as a promising direction for future improvement.

Future work should explore further optimization of the joint network training mechanism, advanced fusion strategies and refined loss formulations. Moreover, extending the framework to incorporate additional multimodal representations (e.g., articulatory or visual information) and validating it on larger-scale EL speaker datasets will be potential directions.

6 Conclusions

6.1 Summary of this Thesis

Sequence-to-sequence (seq2seq) Voice Conversion (VC) exhibits greater potential compared with conventional frame-wise VC, particularly in its ability to flexibly handle the temporal structure from source and target speech, thereby achieving higher naturalness in conversion. Motivated by this advantage, this thesis aimed to employ seq2seq modeling for ElectroLaryngeal-speech-to-normal-SPeech conversion (EL2SP). Applying seq2seq VC to EL2SP entailed three major challenges. The first one was how to alleviate the data-hungry nature of seq2seq models under practically limited EL2SP datasets. The second one was how to enhance robustness against real-world interferences. The third one was how to further enhance the mapping accuracy during training. The core contribution of this thesis lies in proposing and systematically evaluating a series of training techniques that address these challenges while keeping the seq2seq architecture intact. These techniques collectively improve performance, robustness, and data efficiency of seq2seq EL2SP, thereby expanding its feasibility for real-world applications. An overview of these motivations, challenges, and contributions was presented in Chapter 1.

Chapter 2 provided the theoretical context and foundational knowledge as prerequisites for the proposed methods. It introduced the common framework, Transformer architecture, and internal components of seq2seq VC, which were adopted in this thesis. In addition, It surveyed related works including data augmentation, transfer learn-

ing, Speech Enhancement (SE), and multimodal fusion. These aspects offered the basis and inspiration for this thesis, and were rediscussed in subsequent chapters to support the proposed training strategies. It also summarized the evaluation metrics employed throughout this thesis, providing a common foundation for experiments.

The first major contribution of this thesis was to propose pretraining and fine-tuning techniques to enhance the effectiveness of transfer learning under limited data available. Chapter 3 specifically addressed the low-resource data and unideal transferability with a new multi-stage transfer learning framework using data augmentation. Unlike the mainstream approach depending on high-quality Synthetic Data (SD), which are impractical to obtain, the proposed method directly employed low-quality SD as data augmentation. The pretraining began with the knowledge transfer from Text-To-Speech (TTS) to VC owing to a more easily accessible TTS database. Also, TTS fine-tuning was conducted using an original EL2SP dataset to obtain basic EL and normal TTS models for providing low-quality SD and necessary components. To enhance the transferability, a novel encoder adaptation training strategy was introduced, initializing the encoder with the pretrained VC and the decoder with the EL TTS model, thus improving the encoder's comprehending of EL speech. Hereafter, a two-stage fine-tuning scheme was designed that progressively smoothed the transfer learning by combining parallel synthetic and original speech. Since both adaptation and fine-tuning stages leveraged low-quality SD, the entire pipeline was unified and specially practical for the low-resource scenario. Through systematic experiments, the proposed systems consistently outperformed the conventional pretraining–fine-tuning baseline and validated the effectiveness of specific training stages, while offering fundamental insights for the following tasks of EL2SP.

The second major contribution of this thesis was the development of robust training techniques for seq2seq EL2SP. Chapter 4 aimed to address the inability of current seq2seq EL2SP systems to cope with real-world interferences, including noise or reverberation, under limited data. To this end, a many-to-one two-stage fine-tuning

strategy based on SD was proposed to continuously refine the model performance, inspired by the contributions in Chapter 3. Compared with those relying on extra SE modules, the proposed method maintained the single seq2seq architecture and was driven by training data; by injecting various types of interferences, the speech with fine-grained real-word conditions was generated to map with the clean target, finalizing multiple systems for evaluation. Moreover, a series of SE-based baseline systems were constructed and the two-stage training concept was extended to SE modules. Experimental results showed that the two-stage training enhanced SE performance as well, benefiting the entire baseline system. Nevertheless, the proposed systems still outperformed these baselines with its simpler architecture. Analysis of intermediate representations further supported the effectiveness of the proposed method, showing that many-to-one mappings enable seq2seq models to filter out interferences while preserving useful linguistic representations for more accurate conversion.

Recognizing that speech-only modeling could not fundamentally overcome the mapping deviations caused by the significant mismatch between EL and normal speech, the final contribution of this thesis focused on enhancing representation learning with auxiliary information. Chapter 5 proposed a framework that explicitly integrated text and speech representations. A text encoder was first introduced into the seq2seq framework, realizing a joint training framework with dual text and speech inputs to learn richer integrated representations. Then reverting to the seq2seq setting, the reconstruction training was proposed to force the speech encoder to align with the integrated representations, building a better-performing EL2SP system without increasing the complexity. Furthermore, the reconstruction loss was refined by incorporating the integrated representations as an auxiliary loss term, which yielded further improvements in conversion intelligibility. Experimental results confirmed that the proposed method effectively outperformed the speech-representation-based approach, underscoring the benefits of multimodal integration and loss refinement.

In summary, this thesis expanded the ideas of transfer learning and data augmen-

tation by proposing a series of training techniques to efficiently address the practical challenges in EL2SP. At the same time, maintaining a unified architecture further ensured the simplicity and practicality of the proposed methodology.

6.2 Future Work

While the methods proposed in this thesis improved the performance and robustness of seq2seq EL2SP in real-world applications, some limitations remain, along with potential aspects for further exploration. A few outlooks to pursue are provided below.

1. **Toward Framework Generalization:** Chapters 4 and 5 demonstrated the potential of the seq2seq models in handling multi-source inputs. Nonetheless, in Chapter 4, the performance of the proposed method under severe interferences still left much room for improvement compared with that under clean condition. A promising direction is therefore to generalize the proposed methods by integrating the methodologies developed for different challenges. In particular, the framework introduced in Chapter 5, which integrated text and speech representations, may be combined with the robust training approach of Chapter 4. Since text representations have been demonstrated as effective for improving mapping precision, they may also help disentangle useful linguistic information from interfered signals, thereby enhancing model robustness in noisy and/or reverberant conditions. Building on this insight, future work could focus on a more unified framework that simultaneously leverages text inputs and robust fine-tuning strategies to the real-world scenarios. Exploring this integration would also extend the methodological generalization of this thesis.

2. **Extension for multilingual and multi-speaker EL2SP:** Throughout this thesis, the proposed methods focused primarily on single-speaker and monolingual conditions, which limited their applicability in broader scenarios, such as for bilingual EL speakers or systems that must generalize across multiple speakers.

Therefore, a practical future direction is to extend EL2SP to multilingual and multi-speaker dimensions.

- For the multi-speaker dimension, considering recording patient speech is laborious as mentioned in Subsection 1.1.2, a larger database can be constructed by simulating EL speech from additional normal speakers, pairing it with their healthy recordings to form pseudo EL2SP data. Although simulated EL speech differs from real EL speech [11], such data are expected to promote more generalizable representations, particularly when combined with data augmentation.
- For the multilingual dimension, a harder case arises when only one language of a bilingual EL speaker's normal speech (e.g., Japanese) is available, while the other (e.g., English) is impossible to collect. In this situation, leveraging cross-lingual VC [36], [102], [103], [158], [159], [160] or multi-speaker TTS [161]–[163] is a potential direction. For instance, extracting embeddings of other English corpora or text as references, which are then combined with the speaker identity from normal Japanese speech of EL speaker, thus generating pseudo-normal English speech to build the bilingual EL2SP dataset. Furthermore, since bilingual speakers often retain accents when speaking a second language, combining techniques in accent conversion [164]–[166] for the construction of more realistic bilingual EL2SP datasets or customized systems for dialectal conditions could be a more interesting direction. Finally, as pretraining in this thesis relied on a single-speaker TTS corpus, future research could explore pretraining seq2seq models on multi-speaker corpora like the work by Kameoka *et al.* [167], thereby providing stronger, more transferable priors for such a topic.

3. **Extension for experimental validation:** As mentioned earlier, the experi-

mental evaluations in this thesis were conducted under limited speaker coverage due to practical constraints in EL speech data collection. Following the data construction strategies for multi-speaker and multilingual EL2SP discussed above, it becomes feasible to further validate the robustness of the proposed methods at a larger data scale. Thus, one direction for future work is to extend the experimental validation to a wider range of datasets involving more speakers and diverse recording conditions. Evaluating the proposed systems across multiple datasets would allow a more thorough assessment of their robustness and generalization capability.

4. **Extension of subjective evaluation design:** In this thesis, subjective listening tests were conducted following standard protocols in VC research, with evaluations provided by general listeners who are native speakers of the language of the evaluated speech. While this setting allows for consistent and fair comparisons across systems, it does not fully capture the range of perceptual perspectives that may arise in EL2SP scenarios. A promising future direction is to involve a broader range of listener groups, including medical experts such as speech pathologists or clinicians, and family members or frequent communication partners of laryngectomees. These listener groups may exhibit different perceptual sensitivities or evaluation criteria compared with general listeners. Investigating potential systematic differences or biases across listener groups would provide deeper insights into the perceptual evaluation of EL2SP systems.

5. **Development toward low-latency EL2SP system:** To further facilitate daily convenience for laryngectomees, a potential real-world application is to deploy the proposed methods on mobile platforms, such as smartphones or tablets, thereby achieving more efficient communication. Consequently, developing EL2SP systems with lower-latency, real-time capability is a research direction worth exploring. The proposed EL2SP systems in this thesis are based on an autoregres-

sive seq2seq architecture with approximately 30.4 million trainable parameters. On a single GPU, the end-to-end inference time for a 4–5 second utterance is approximately 0.87 seconds for the seq2seq model, with an additional 0.04 seconds for vocoder synthesis. This performance is acceptable for some offline or assistive applications, for instance through server-side (e.g., cloud-based) inference using GPU resources, where input speech is processed on a backend and then returned to the device. However, such a deployment paradigm would inevitably introduce additional communication latency. As a result, the current latency and computational complexity still pose challenges for real-time deployment on resource-constrained wearable devices.

Due to the complexity of seq2seq modeling, previous works on low-latency VC mostly focused on frame-wise architectures [26], [168]. Although these methods achieve relatively efficient conversion, they still suffer from suboptimal conversion intelligibility and naturalness, which are crucial metrics for EL2SP. In recent years, non-autoregressive seq2seq models such as FastSpeech2 TTS [169] and VC [170], have provided greater potential. Furthermore, in EL2SP, study by Kobayashi *et al.* [171] introduced the design concept of FastSpeech2 into the Convolutional Long-short term Deep Neural Network (CLDNN) [2] modules, extending frame-wise alignment to seq2seq modeling. Concretely, by incorporating a length regulator and variance adaptors with multi-factor predictors under a non-autoregressive structure, this method improved the conversion naturalness while effectively accelerating inference, reaching an overall latency to a practical range of around 51 ms. This progress provides valuable insights for the future research and application of low-latency deployment of the proposed systems in this thesis.

6. **Improvement of multimodal architectures:** Recall that the text–speech integration method in Chapter 5 showed encouraging results, demonstrating ben-

eficial from learning integrated modalities for EL2SP. This suggests that exploring additional modalities and more advanced integration/alignment strategies represents promising research directions such as the following:

- From the modality usage perspective, integrating multimodal features to improve semantic integrity and speech quality is promising, as reflected in Chapter 2. In this light, combining text and visual modalities is expected to provide clearer linguistic and contextual cues than using text alone, holding greater potential for enhancing both mapping precision and conversion robustness in real-world EL2SP application. To this end, constructing multimodal EL2SP databases that incorporate audio, text, and visual clips will be an essential prerequisite. Another limitation of current EL2SP datasets lies in the absence of expressive elements such as prosody and emotional nuance in EL speech, while normal target speech also cannot fully cover the variety of affective expressions required in real-world communication. Inspired by some works [133]–[135], an interesting subtask would be to integrate emotion or facial expression labels, or leverage prompt-based techniques to enrich the emotional semantics, thereby enabling EL speech to be converted into natural speech with emotional styles, which is more practical for interactions.

- From the methodological perspective, future work should also focus on enhancing the training algorithms of the multimodal framework beyond the current reconstruction-based design. At present, incorporated multimodal features are attended to as a loss factor with a preset weight. This could be extended by adopting trainable weighting schemes and more comprehensive loss objectives that better balances multimodal contributions. In addition, introducing contrastive learning frameworks [172], [173] presents another promising avenue. For instance, inspired by the recent contrastive

text–audio approach [174]–[176], we may speculate that large-scale text–audio contrastive pretraining could be helpful to extract more precise representations for downstream EL2SP. Furthermore, drawing on preliminary insights by Wang *et al.* [130], systematically investigating the optimal positions and approaches for modality fusion in terms of the dimension of information input within the current EL2SP framework, is an another promising research direction.

7. **Further investigation on different model architectures:** The proposed methods in this thesis were primarily developed within a Transformer-based framework, which provided a consistent setting to validate the effectiveness of different training strategies. A natural extension is to investigate the applicability of these strategies to more recent generative architectures, such as diffusion-based [177] and flow-based [178] models, which show strong potential for high-fidelity speech generation. For instance, Hsu *et al.* [179] and Chen *et al.* [180] applied diffusion-based VC models to dysarthric speech enhancement, demonstrating encouraging results particularly in speaker identity reconstruction. Inspired by these efforts, exploring diffusion-based EL2SP may offer an alternative formulation. In addition, Lu *et al.* [181] demonstrated the effectiveness of diffusion models for speech enhancement under noisy conditions, including improved generalization to unseen noise data. This finding suggests a potential direction for further improving the interference-robust EL2SP systems of Chapter 4. On the other hand, Proszewska *et al.* [182] built a normalizing-flow model in the mel-spectrogram space and explicitly disentangles content, pitch, and speaker factors for language-independent, text-free VC. It reported substantially improved intelligibility, even for languages unseen during training. Such robustness indicates that flow-based generation mechanisms may provide useful insights for future EL2SP research, particularly in multilingual and multi-speaker settings.

Taken collectively, addressing these open challenges and exploring more comprehensive frameworks in future work will not only further enhance the performance of seq2seq EL2SP systems but also broaden their applicability to more diverse and realistic scenarios.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Tomoki Toda, for his invaluable guidance, unwavering support, and profound wisdom throughout my doctoral journey. This thesis would not have been possible without his mentorship. Since our first contact in 2019, when I was a novice with a limited knowledge in Voice Conversion and Speech Synthesis, Prof. Toda has continuously encouraged me and offered me the opportunity to join his laboratory. Looking back, it is difficult to imagine how much progress I have made, and all of it is built upon his support. His rigor, intelligence, persistence, rationality have deeply influenced me, not only as a researcher but also as a person. He has always been a role model of what I envision as a truly outstanding scholar. Beyond academic guidance, he has also provided financial support that enabled me to concentrate fully on my research, as well as life lessons through his self-discipline. For example, his habit of running daily even during overseas business trips and updating his new records at each year's kick-off meeting. These moments inspired me to cultivate more discipline in my life. I could not be more fortunate to have privilege of being mentored by him, and I sincerely wish him and his family all the best in the future.

I sincerely thank Prof. Ichiro Ide and Prof. Yu Tsao for serving on my thesis review committee, and for their valuable time, sustained commitment, and constructive comments during the entire review process, which greatly improved this thesis.

I am also deeply grateful to Dr. Kazuhiro Kobayashi for his insightful guidance on my research in electrolaryngeal speech enhancement. At our weekly doctoral meet-

ings, his constructive advice always helped me clear confusion, and the dataset he built became an essential resource for my work. My achievements owe much to his dedicated support.

My heartfelt thanks also go to Prof. Wen-chin Huang, whose encouragement and guidance have been invaluable since the very beginning of my PhD journey. When I first joined the lab, he kindly reached out to me and offered me a project. His sharp insights and consistent follow-up helped me overcome the clumsiness of a beginner. Our first collaboration even received a prize, which was an unforgettable honor for me. I would not have entered this field so smoothly without his support. Since then, he has continued to provide me with constructive advice, for which I am truly grateful.

I would like to thank my collaborator and friend Mr. Lester Violeta for his intelligence and passion. Working with him enabled me to complete important journal works. Sharing conference experiences and life stories, and enjoying food together, including spicy Sichuan dishes and hotpot, have all been wonderful memories. My sincere gratitude extends to Mr. Takuya Fujimura, Mr. Yeonjong Choi, and Mr. Chao Xie for their support on speech enhancement. Choi-san and Xie-san shared valuable ideas while Fujimura-san kindly provided his toolkit, facilitating the progress of my projects and papers. I am also grateful to Dr. Fengji Li for collaboration in Biomedical Engineering during his exchange in NU, as well as the memories of our time together in Copenhagen and Orlando, including our adventures at Disney World and Universal Studios.

My heartfelt thanks go to Ms. Nami Noro for her professionalism, kindness, and constant support. She always handled administrative matters in advance, making my research life and conference travels much smoother. Her warm concern for my daily life has been a source of comfort.

I also sincerely thank my laboratory colleagues and friends, Dr. Shaowen Chen, Dr. Shuming Luan, Dr. Rui Wang, Mr. Jiajun He, and Mr. Xiaohan Shi, for the memories and friendships we built —cycling under sunshine and moonlight, celebrating

birthdays, and sharing meals.

Even those who have left the laboratory still shine brightly in my memory. Recall early days of my PhD, the lab was often only Dr. Patrick Lumban Tobing and me due to the pandemic. We shared enjoyable research moments and he gave me generous advice and guidance, for which I am sincerely grateful. I also appreciate Ms. Mayuko Hayashi for her much support, including helping me with residence registration and caring for my well-being. Her kindness gave me much comfort during my early days in Japan. I would equally thank to Mr. Tatsunori Uchino, who helped me with hardware/software issues and even made calls to the bank when my Japanese was still poor.

I also owe special thanks to my long-time friends. Dr. Huang (Arizona State University), thank you for years of friendship and advices across many aspects. Dr. Xu, thank you for your long-standing encouragement for both academically and personally, and for visiting me in Japan. Mr. Cui, thank you for decades of friendship and for flying across the U.S. to meet me in Florida. I am also grateful to my former roommate, Mr. Li, with whom I often stayed up late discussing technical problems —those conversations were inspiring and unforgettable. My deep gratitude also goes to Mr. Xu, Miss Xu, Mr. Li, and Mr. Zhang for their constant care throughout my years abroad.

I owe my deepest gratitude to my family. Thank you to my parents for their endless patience, wise counsel, and unconditional support, especially during my time studying abroad. I also want to express my heartfelt thanks to my girlfriend, my research companion, for her love and support throughout this journey. Thank you for standing by me with constant encouragement and understanding. I am deeply grateful to have shared this journey with you, and I look forward to many wonderful memories together in the future.

There are many more people whose names I could not list here, but I am sincerely thankful to everyone who has supported me in any way during this journey.

Bibliography

[1] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[2] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN," in *Proceedings of 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2115–2119.

[3] E. Babin, D. Beynier, D. Le Gall, and M. Hitier, "Psychosocial quality of life in patients after total laryngectomy," *Revue de Laryngologie-Otologie-Rhinologie*, vol. 130, no. 1, pp. 29–34, 2009.

[4] B. Polat, K. S. Orhan, M. C. Kesimli, Y. Gorgulu, M. Ulusan, and K. Deger, "The effects of indwelling voice prosthesis on the quality of life, depressive symptoms, and self-esteem in patients with total laryngectomy," *European Archives of Oto-Rhino-Laryngology*, vol. 272, no. 11, pp. 3431–3437, 2015.

[5] C. G. Tang and C. F. Sinclair, "Voice restoration after total laryngectomy," *Otolaryngologic Clinics of North America*, vol. 48, no. 4, pp. 687–702, 2015.

[6] M. I. Singer and E. D. Blom, "An endoscopic technique for restoration of voice after laryngectomy," *Annals of Otology, Rhinology & Laryngology*, vol. 89, no. 6, pp. 529–533, 1980.

[7] M. Hashiba, N. Vemi, M. Oikawa, Y. Yamaguchi, Y. Sugai, and T. Ifukube, "Industrialization of the electrolarynx with a pitch control function and its evalu-

ation," *IEICE Transactions Information and Systems (Japanese Edition)*, vol. J84-D2, no. 6, pp. 1240–1247, 2001.

[8] S. E. Williams and J. B. Watson, "Differences in speaking proficiencies in three laryngectomee groups," *Archives of Otolaryngology*, vol. 111, no. 4, pp. 216–219, 1985.

[9] K. Ma, P. Demirel, C. Y. Espy-Wilson, and J. MacAuslan, "Improvement of electrolaryngeal speech by introducing normal excitation information," in *Proceedings of 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999, pp. 323–326.

[10] P. Stanislav, J. V. Psutka, and J. Psutka, "Recognition of the electrolaryngeal speech: Comparison between human and machine," in *Proceedings of 20th International Conference on Text, Speech, and Dialogue*, 2017, pp. 509–517.

[11] L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Pretraining and adaptation techniques for electrolaryngeal speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2777–2789, 2024.

[12] D. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1985*, 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] 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.

[15] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[16] K. Ito and L. Johnson, *The LJ Speech Dataset*, https://keithito.com/LJ-Speech-Dataset/, 2017. Accessed: 2025-12-15.

[17] V. Christophe, Y. Junichi, M. Kirsten, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*, https://datashare.ed.ac.uk/handle/10283/2950, *University of Edinburgh, The Centre for Speech Technology Research (CSTR)*, 2016. Accessed: 2025-12-15.

[18] S. Ando and H. Fujihara, "Construction of a large-scale Japanese ASR corpus on TV recordings," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*, 2021, pp. 6948–6952.

[19] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proceedings of 5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[20] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Proceedings of 17th Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 1632–1636.

[21] 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 *Proceedings of Odyssey 2018 the Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[22] 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 *Proceedings of ISCA Joint Workshop for Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[23] 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.

[24] 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.

[25] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 172–183, 2013.

[26] K. Kobayashi and T. Toda, "Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN," in *Proceedings of 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 396–400.

[27] Z. Qian, H. Niu, L. Wang, K. Kobayashi, S. Zhang, and T. Toda, "Mandarin electro-laryngeal speech enhancement based on statistical voice conversion and manual tone control," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2021*, 2021, pp. 546–552.

[28] Z. Cai, Z. Xu, and M. Li, "F0 contour estimation using phonetic feature in electrolaryngeal speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 2019, pp. 6490–6494.

[29] 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. E99-D, no. 7, pp. 1877–1884, 2016.

[30] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[31] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Proceedings of 9th Annual Conference of the International Speech Communication Association (Interspeech)*, 2008, pp. 1453–1456.

[32] G. Kotani, H. Suda, D. Saito, and N. Minematsu, "Experimental investigation on the efficacy of affine-DTW in the quality of voice conversion," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2019*, 2019, pp. 119–124.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3104–3112, 2014.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of 3rd International Conference on Learning Representations*, 2015, 15 pages.

[35] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of Conference on Empirical Methods in Natural Language Processing 2014*, 2014, pp. 1724–1734.

[36] 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 *Proceedings of ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 160–164.

[37] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*, 2021, pp. 5944–5948.

[38] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-Sequence Voice Conversion with Attention and context preservation mechanisms," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 2019, pp. 6805–6809.

[39] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," *arXiv Preprint*, arXiv:1704.02360, 2017.

[40] M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. R. Jang, and H.-M. Wang, "Mandarin electrolaryngeal speech voice conversion with sequence-to-sequence modeling," in *Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 650–657.

[41] W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022*, 2022, pp. 6672–6676.

[42] 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 *Proceedings of 22nd Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 1329–1333.

[43] 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 *Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 4686–4690.

[44] L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Intermediate fine-tuning using imperfect synthetic speech for improving electrolaryngeal speech recognition,"

in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, 2023, pp. 1–5.

[45] D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Two-stage training method for Japanese electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion," in *Proceedings of 2023 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 949–954.

[46] L. P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, and T. Toda, "Electrolaryngeal speech intelligibility enhancement through robust linguistic encoders," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024*, 2024, pp. 10 961–10 965.

[47] 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," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 2019, pp. 6785–6789.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[50] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[51] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[52] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image Transformers," *arXiv Preprint*, arXiv:2106.08254, 2021.

[53] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi, "Big self-supervised models advance medical image classification," in *Proceedings of 18th IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3478–3488.

[54] J. Fu, S. Xu, H. Liu, Y. Liu, N. Xie, C.-C. Wang, J. Liu, Y. Sun, and B. Wang, "CMA-CLIP: Cross-Modality Attention CLIP for text-image classification," in *Proceedings of IEEE International Conference on Image Processing (ICIP) 2022*, 2022, pp. 2846–2850.

[55] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEiT pretraining for all vision and vision–language tasks," *arXiv Preprint*, arXiv:2208.10442, 2022.

[56] M. Xu, M. Islam, C. M. Lim, and H. Ren, "Class-incremental domain adaptation with smoothing and calibration for surgical report generation," in *Proceedings of 24th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Part III*, 2021, pp. 269–278.

[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *Proceedings of 2019 North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[58] A. Radford, K. Narasimahan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pre-training*, OpenAI preprint, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. June 11, 2018. Accessed: 2025-12-15.

[59] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, 24 pages, Feburary 14, 2019.

[60] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[61] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report," *arXiv Preprint*, arXiv:2303.08774, 2023.

[62] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "Wav2vec: Unsupervised pre-training for speech recognition," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 3465–3469.

[63] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, vol. 32, pp. 3171–3180, 2019.

[64] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proceed-*

ings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018, 2018, pp. 4889–4893.

[65] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[66] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proceedings of ISCA Workshop on Speech Synthesis (SSW 9)*, 2016, pp. 146–152.

[67] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for phase-aware speech enhancement," *arXiv Preprint*, arXiv:2008.00264, 2020.

[68] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Proceedings of 19th Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 342–346.

[69] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy voice conversion framework with denoising model," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2021*, 2021, pp. 814–820.

[70] Y. Choi, C. Xie, and T. Toda, "An evaluation of three-stage voice conversion framework for noisy and reverberant conditions," in *Proceedings of 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 4910–4914.

[71] S. Yang, Y. Wang, and L. Xie, "Adversarial feature learning and unsupervised clustering based speech synthesis for found data with acoustic and textual noise," *IEEE Signal Processing Letters*, vol. 27, pp. 1730–1734, 2020.

[72] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 4115–4119.

[73] 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," in *Proceedings of ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 165–169.

[74] 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 *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 870–877.

[75] C.-Y. Huang, K.-W. Chang, and H.-Y. Lee, "Toward degradation-robust voice conversion," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022*, 2022, pp. 6777–6781.

[76] Y. Choi, C. Xie, and T. Toda, "Reverberation-controllable voice conversion using reverberation time estimator," in *Proceedings of 24th Annual Conference of the International Speech Communication Association (Interspeech)*, 2023, pp. 2103–2107.

[77] O. Slizovskaia, J. Janer, P. Chandna, and O. Mayor, "Voice conversion with limited data and limitless data augmentations," *arXiv Preprint*, arXiv:2212.13581, 2022.

[78] D. Ma, Y. Choi, F. Li, C. Xie, K. Koboyashi, and T. Toda, "Robust sequence-to-sequence voice conversion for electrolaryngeal speech enhancement in noisy and reverberant conditions," in *Proceedings of Annual International Conference*

of 46th IEEE Engineering in Medicine and Biology Society (IEEE EMBC), 2024, 4 pages.

[79] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source Tacotron and WaveNet," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 15–19.

[80] H.-T. Luong and J. Yamagishi, "Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech," in *Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 200–207.

[81] S.-w. Park, D.-y. Kim, and M.-c. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data," in *Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 4696–4700.

[82] 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 *Proceedings of 22nd Annual Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4676–4680.

[83] 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 *Proceedings of 18th Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 4006–4010.

[84] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,"

in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, 2018, pp. 4779–4783.

[85] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

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

[87] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[88] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with Transformer network," in *Proceedings of 33rd AAAI Conference on Artificial Intelligence*, 2019, pp. 6706–6713.

[89] 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 *Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 449–456.

[90] K. Ezzine, J. Di Martino, and M. Frikha, "Intelligibility improvement of esophageal speech using sequence-to-sequence voice conversion with auditory attention," *Applied Sciences*, vol. 12, no. 14, pp. 7062–7077, 2022.

[91] Y. Yang, H. Zhang, Z. Cai, Y. Shi, M. Li, D. Zhang, X. Ding, J. Deng, and J. Wang, "Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion," *Biomedical Signal Processing and Control*, vol. 80, pp. 104 279–104 310, 2023.

[92] L. Dong, S. Xu, and B. Xu, "Speech-Transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, 2018, pp. 5884–5888.

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

[94] R. Liu, X. Chen, and X. Wen, "Voice conversion with Transformer network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 2020, pp. 7759–7759.

[95] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, 2018, pp. 4784–4788.

[96] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 2020, pp. 7654–7658.

[97] M. Baas and H. Kamper, "Voice conversion can improve ASR in very low-resource settings," in *Proceedings of 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 3513–3517.

[98] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*, 2021, pp. 6593–6597.

[99] B. Chen, Z. Xu, and K. Yu, "Data augmentation based non-parallel voice conversion with frame-level speaker disentangler," *Speech Communication*, vol. 136, pp. 14–22, 2022.

[100] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of 35th International Conference on Machine Learning*, 2018, pp. 3918–3926.

[101] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 2020, pp. 7744–7748.

[102] M. Zhang, Y. Zhou, L. Zhao, and H. Li, "Transfer learning from speech synthesis to voice conversion with non-parallel training data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1290–1302, 2021.

[103] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 2019, pp. 6790–6794.

[104] X. Tian, E. S. Chng, and H. Li, "A speaker-dependent WaveNet for voice conversion with non-parallel data," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 201–205.

[105] S. Liu, Y. Cao, and H. Meng, "Multi-target emotional voice conversion with neural vocoders," *arXiv Preprint*, arXiv:2004.03782, 2020.

[106] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[107] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and 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.

[108] C. V. Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proceedings of 17th Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 352–356.

[109] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 2019, pp. 5901–5905.

[110] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2000.

[111] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proceedings of 4th IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 313–317.

[112] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 6, pp. 1411–1418, 2014.

[113] R. Aihara, T. Fujii, T. Nakashika, T. Takiguchi, and Y. Ariki, "Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 32–41, 2015.

[114] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, pp. 151–164, 2019.

[115] B. Van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the Zerospeech 2020 Challenge," in *Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 4836–4840.

[116] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022*, 2022, pp. 6787–6791.

[117] C. Xie and T. Toda, "Noisy-to-noisy voice conversion under variations of noisy condition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3871–3882, 2023.

[118] Y.-J. Chan, C.-J. Peng, S.-S. Wang, H.-M. Wang, Y. Tsao, and T.-S. Chi, "Speech enhancement-assisted StarGan voice conversion in noisy environments," *arXiv Preprint*, arXiv:2110.09923, 2021.

[119] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[120] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, and W. Post, "The AMI meeting corpus: A pre-announcement," in

Proceedings of 2nd International Workshop on Machine Learning for Multimodal Interaction (MLMI), 2005, pp. 28–39.

[121] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," *arXiv Preprint*, arXiv:1803.10609, 2018.

[122] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in *Proceedings of 17th Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 3743–3747.

[123] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1876–1887, 2017.

[124] J.-C. Chou, C.-C. Yeh, and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 664–668.

[125] H. Du, L. Xie, and H. Li, "Noise-robust voice conversion with domain adversarial training," *Neural Networks*, vol. 148, pp. 74–84, 2022.

[126] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *Proceedings of 8th IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 552–559.

[127] A. Mottini, J. Lorenzo-Trueba, S. V. K. Karlapati, and T. Drugman, "Voicy: Zero-shot non-parallel voice conversion in noisy reverberant environments," in *Proceedings of 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 113–117.

[128] Y. Choi, C. Xie, and T. Toda, "Reverberation-controllable voice conversion using reverberation time estimator," in *Proceedings of 24th Annual Conference of*

the International Speech Communication Association (Interspeech), 2023, pp. 2103–2107.

[129] S. Ghosh, U. Tyagi, S. Ramaneswaran, H. Srivastava, and D. Manocha, "MMER: Multimodal Multi-task learning for speech Emotion Recognition," *arXiv Preprint*, arXiv:2203.16794, 2022.

[130] Y. Wang, Y. Gu, Y. Yin, Y. Han, H. Zhang, S. Wang, C. Li, and D. Quan, "Multimodal Transformer augmented fusion for speech emotion recognition," *Frontiers in Neurorobotics*, vol. 17, no. 1181598, pp. 1–10, 2023.

[131] W. Fan, X. Xing, B. Cai, and X. Xu, "MGAT: Multi-Granularity Attention based Transformers for multi-modal emotion recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, 2023, 5 pages.

[132] Z. Chen, X. Li, Z. Ai, and S. Xu, "StyleFusion TTS: Multimodal Style-control and enhanced feature Fusion for zero-shot Text-To-Speech synthesis," in *Proceedings of 7th Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2024, pp. 263–277.

[133] W. Guan, Y. Li, T. Li, H. Huang, F. Wang, J. Lin, L. Huang, L. Li, and Q. Hong, "MM-TTS: Multi-Modal prompt based style transfer for expressive Text-To-Speech synthesis," in *Proceedings of 38th AAAI Conference on Artificial Intelligence*, 2024, pp. 18 117–18 125.

[134] H. Kameoka, K. Tanaka, A. V. Puche, Y. Ohishi, and T. Kaneko, "Crossmodal voice conversion," *arXiv Preprint*, arXiv:1904.04540, 2019.

[135] X. Niu, J. Zhang, and C. P. Martin, "HybridVC: Efficient voice style conversion with text and audio prompts," *arXiv Preprint*, arXiv:2404.15637, 2024.

[136] R. Li, R. Huang, L. Zhang, J. Liu, and Z. Zhao, "AlignSTS: Speech-To-Singing conversion via cross-modal Alignment," *arXiv Preprint*, arXiv:2305.04476, 2023.

[137] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.

[138] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR —Half-baked or well done?" In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 2019, pp. 626–630.

[139] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2010*, 2010, pp. 4214–4217.

[140] International Telecommunication Union Telecommunication Standardization Sector, *Methods for Subjective Determination of Transmission Quality*. International Telecommunication Union, 1996.

[141] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv Preprint*, arXiv:1711.00354, 2017.

[142] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end Speech Processing toolkit," in *Proceedings of 19th Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 2207–2211.

[143] 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 *Proceedings of 8th International Conference on Learning Representations*, 2020, 36 pages.

[144] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 2020, pp. 6199–6203.

[145] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 5036–5040.

[146] L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition," in *Proceedings of 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 41–45.

[147] W.-C. Huang, S.-W. Yang, T. Hayashi, and 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.

[148] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of 38th International Conference on Machine Learning*, 2021, pp. 5530–5540.

[149] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 2020, pp. 696–700.

[150] T. Fujimura and T. Toda, "Analysis of noisy-target training for DNN-based speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, 2023, 5 pages.

[151] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015*, 2015, pp. 5206–5210.

[152] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2015*, 2015, pp. 504–511.

[153] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv Preprint*, arXiv:2110.07840, 2021.

[154] M. Kawanaka, Y. Koizumi, R. Miyazaki, and K. Yatabe, "Stable training of DNN for speech enhancement based on perceptually-motivated black-box cost function," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 2020, pp. 7524–7528.

[155] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," in *Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 2637–2641.

[156] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[157] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for dimension reduction," *arXiv Preprint*, arXiv:1802.03426, 2018.

[158] M. Zhang, Y. Zhou, Y. Ren, C. Zhang, X. Yin, and H. Li, "RefXVC: Cross-lingual Voice Conversion with enhanced Reference leveraging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4146–4156, 2024.

[159] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 674–678.

[160] P. Lumban Tobing, Y.-C. Wu, and T. Toda, "Baseline system of Voice Conversion Challenge 2020 with cyclic variational autoencoder and Parallel WaveGAN," in *Proceedings of Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 155–159.

[161] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2080–2084.

[162] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a mixed-lingual neural TTS system with only monolingual data," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2060–2064.

[163] R. Badlani, R. Valle, K. J. Shih, J. F. Santos, S. Gururani, and B. Catanzaro, "Multilingual multiaccented multispeaker TTS with RADTTS," *arXiv Preprint*, arXiv:2301.10335, 2023.

[164] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, 2018, pp. 5314–5318.

[165] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2843–2847.

[166] T.-N. Nguyen, N.-Q. Pham, and A. Waibel, "Accent conversion using pre-trained model and synthesized data from voice conversion," in *Proceedings of 23rd Annual Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 2583–2587.

[167] H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda, "Many-to-many voice Transformer network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 656–670, 2020.

[168] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, "Real-time, full-band, online DNN-based voice conversion system using a single CPU," in *Proceedings of 21st Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 1021–1022.

[169] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," *arXiv Preprint*, arXiv:2006.04558, 2020.

[170] H. Kameoka, K. Tanaka, and T. Kaneko, "FastS2S-VC: Streaming non-autoregressive Sequence-to-Sequence Voice Conversion," *arXiv Preprint*, arXiv:2104.06900, 2021.

[171] K. Kobayashi, T. Hayashi, and T. Toda, "Low-latency electrolaryngeal speech enhancement based on FastSpeech2-based voice conversion and self-supervised speech representation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, 2023, 5 pages.

[172] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proceedings of IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023, 2023, 5 pages.

[173] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language–audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 5 pages.

[174] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024*, 2024, pp. 336–340.

[175] C. Qiang, H. Li, Y. Tian, R. Fu, T. Wang, L. Wang, and J. Dang, "Learning speech representation from contrastive token–acoustic pretraining," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024*, 2024, pp. 10 196–10 200.

[176] Z. Wu, Q. Li, S. Liu, and Q. Yang, "DCTTS: Discrete diffusion model with Contrastive learning for Text-To-Speech generation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024*, 2024, pp. 11 336–11 340.

[177] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for Text-To-Speech," in *Proceedings of 38th International Conference on Machine Learning*, PMLR, 2021, pp. 8599–8608.

[178] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative flow for Text-To-Speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[179] W.-S. Hsu, G.-T. Lin, and W.-H. Wang, "Enhancing dysarthric voice conversion with fuzzy expectation maximization in diffusion models for phoneme prediction," *Diagnostics*, vol. 14, no. 23, pp. 1–18, 2024.

[180] X. Chen, D. Yang, W. Wu, M. Wu, J. Xu, X. Wu, Z. Wu, and H. Meng, "DiffDSR: Dysarthric Speech Reconstruction using latent Diffusion model," in *Proceedings of 26th Annual Conference of the International Speech Communication Association (Interspeech)*, 2025, pp. 2113–2117.

[181] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022*, 2022, pp. 7402–7406.

[182] M. Proszewska, G. Beringer, D. Sáez-Trigueros, T. Merritt, A. Ezzerg, and R. Barra-Chicote, "GlowVC: Mel-spectrogram space disentangling model for language-independent text-free Voice Conversion," in *Proceedings of 23rd Annual Conference of the International Speech Communication Association (Interspeech) 2022*, 2022, 2973–2977.

[183] Munich Artificial Intelligence Laboratories GmbH, *The M-AILABS Speech Dataset*, https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/, Accessed: Nov. 30, 2019, 2019.

A Appendix

This appendix supplements Chapter 2 by providing detailed experimental investigations. The study systematically examines the impact of Text-To-Speech (TTS)-based Synthetic Parallel Data (SPD) on sequence-to-sequence (seq2seq) Voice Conversion (VC) under non-parallel setting. Some factors, such as a comparison of SPD by source or target speaker, the effects of SPD in a semi-parallel setting including a parallel subset, and the usage of SPD with external text data are investigated. The following questions are addressed:

- **Q1:** What are the feasibility and properties of using SPD?
- **Q2:** How can this method benefit from a semi-parallel setting?
- **Q3:** What are the influences of using external text data?

A.1 Experimental Evaluations For Synthetic Parallel Data (SPD) Effect Investigation

The section is divided into five subsections. Subsection A.1.1 introduces the overall experimental data configuration, and the next three Subsections A.1.2, A.1.3, and A.1.4 address each of the three questions Q1, Q2, and Q3, respectively. Each subsection gives viewpoints and discussions based on results of objective evaluations. Finally, Subsection A.1.5 presents discussions on these three questions based on results of subjective evaluations.

A.1.1 Datasets and Configuration

In this study, the original datasets were from the CMU ARCTIC database [19], containing 1,132 parallel utterances recorded by several English speakers in 16 kHz, where the female speakers named clb and slt, and male speakers named bdl and rms were selected as source or target speakers. Here, 100 utterances from the database were used as the development set and another 100 utterances as an evaluation set, respectively, and then, the remaining 932 utterances were used as the training set. For the external data, an English corpus from the M-AILABS database [183] was chosen with a total of 15,369 utterances, roughly 30 hours long.

The implementations of the open-source ESPnet toolkit [96], [142] was followed, where 80-dimensional mel filterbanks with 1,024 Fast Fourier Transformation (FFT) points and a 256-point frame shift were used to extract the acoustic features. The Transformer-TTS architecture [88] was used as a TTS model to generate SPD. As for the VC model, Voice Transformer Network (VTN) model was directly used. These two models were both pretrained using M-AILABS. To generate high-quality synthetic data, the Parallel WaveGAN (PWG) neural vocoder [144], [160] was used to implement on the TTS and VTN models. In terms of the different target speakers, the corresponding speaker-dependent PWG vocoders were trained by using the full dataset from CMU ARCTIC.

The objective evaluations were performed using several metrics, such as Mel-Cepstrum Distortion (MCD) to capture spectral envelope distortion between generated speech and natural speech, and Word Error Rate (WER) and the Character Error Rate (CER) to evaluate the intelligibility. The ASR engine was a Transformer-based model [92] which was trained by LibriSpeech database [151].

Subjective tests were also conducted to evaluate VC perceptual performance from two perspective: naturalness and speaker similarity of the generated speech. In the naturalness test, an opinion test was conducted to evaluate the naturalness with a

Mean Opinion Score (MOS). Listeners were asked to rate (in 1-to-5 scale) the naturalness of each given speech. In the speaker similarity test, a preference test was conducted. A pair of the converted speech and the ground-truth speech were presented to the listeners at the same time, who were asked to judge whether the two utterances produced by the same speaker on a four-point scale.

A.1.2 Investigation of Feasibility and Property on SPD

This subsection mainly focuses on the experimental investigation of Q1. Q1 is further divided into two subquestions:

- **Q1-1:** How does quality of data affect VC performance?
- **Q1-2:** Which kind of the training pair is better?

For Q1-1, given a certain amount of non-parallel original training data, the TTS models need to be trained in advance using the non-parallel dataset to generate SPD. Here, let us use the term *datasize* to represent the volume of utterances used. Therefore, the original training datasize will affect the TTS model performance and determine the total VC training volume, which can be taken as the two indicators of Q1-1.

For Q1-2, adding SPD will generate different types of training pairs. Hence, different training pairs may affect the VC results, which needs to be further investigated. Here, under the condition that the development set (100) and the evaluation set (100) remain unchanged, the other utterances (932) of the CMU ARCTIC database [19] were chosen to divide different quantities of non-parallel training sets for comparison experiments. The experimental process is shown in Figure A.1. Each experiment is implemented with five groups of source–target combinations as follows:

1. <source natural, target natural> (natural–natural)
2. <source natural, target synthetic> (natural–synthetic)
3. <source synthetic, target natural> (synthetic–natural)
4. <source synthetic, target synthetic> (synthetic–synthetic)

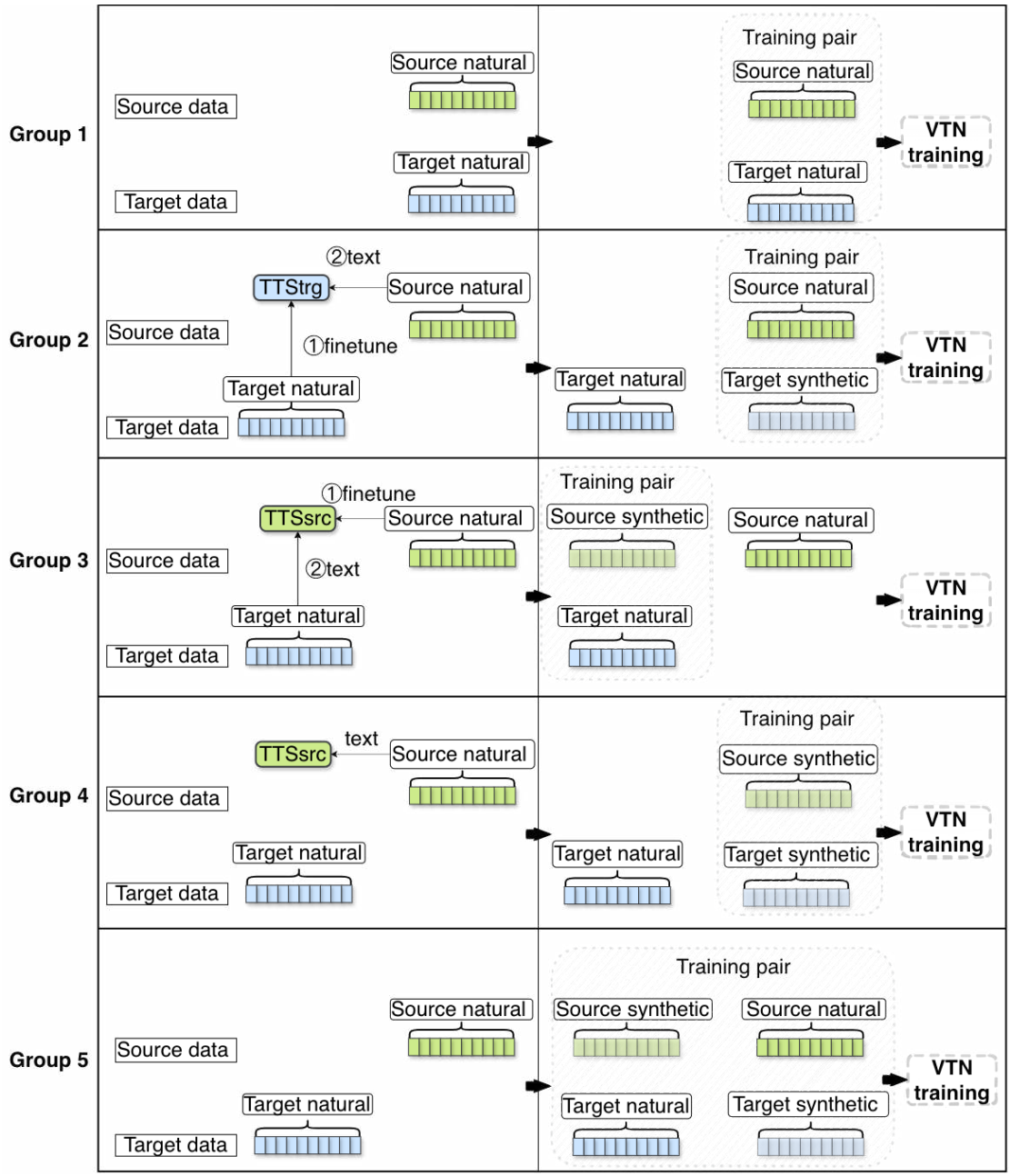


Figure A.1: Implementation of five groups in the experiments. Note that the proportion of synthetic and natural corpora is the same in both source and target speakers in Group 5.

Table A.1: Comparison results with different training pairs and datasizes. TTS-450, TTS-400, TTS-200 and TTS-80 represent the homologous datasize of TTS finetuning, which also reflect TTS performance, the datasize of SPD generation and VC training. "n–n", "n–s", "s–n", "s–s", and "(s + n)–(n + s)" denote natural–natural, natural–synthetic, synthetic–natural, synthetic–synthetic, and (synthetic + natural)–(natural + synthetic), respectively. Note that the VC training datasize of the group (synthetic + natural)–(natural + synthetic) is twice that of the other four groups respectively.

Speaker	Training pair	TTS-450	TTS-400	TTS-200	TTS-80
Source–Target	Description	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]
clb–slt	n–n	6.23 / 2.3 / 4.9	6.35 / 5.0 / 9.1	6.66 / 5.3 / 9.5	6.87 / 4.1 / 8.8
	n–s	**6.72 / 3.3 / 6.4**	**6.64 / 3.7 / 7.5**	**6.74 / 5.6 / 10.7**	**7.27 / 8.3 / 13.1**
	s–n	**6.74 / 4.3 / 8.0**	**6.68 / 3.4 / 6.7**	**6.68 / 3.1 / 6.5**	**6.96 / 5.7 / 11.5**
	s–s	6.77 / 5.3 / 8.2	6.85 / 4.8 / 8.5	6.97 / 8.5 / 13.1	8.33 / 19.6 / 25.6
	(s + n)–(n + s)	6.61 / 4.3 / 8.5	6.59 / 3.7 / 7.1	6.70 / 4.9 / 8.4	7.03 / 8.0 / 12.6
bdl–rms	n–n	6.56 / 7.8 / 14.0	6.64 / 10.3 / 18.7	7.17 / 11.7 / 20.7	7.02 / 16.6 / 27.2
	n–s	**7.02 / 11.5 / 20.7**	**7.32 / 9.7 / 17.1**	**7.43 / 11.3 / 19.5**	**7.72 / 12.7 / 20.4**
	s–n	**6.63 / 10.2 / 18.5**	**6.81 / 9.4 / 16.4**	**6.94 / 10.7 / 18.6**	**7.19 / 11.4 / 18.8**
	s–s	7.07 / 10.2 / 18.8	7.36 / 8.1 / 15.3	7.51 / 11.6 / 20.3	7.82 / 14.4 / 24.8
	(s + n)–(n + s)	6.91 / 8.8 / 16.6	7.29 / 8.9 / 15.0	7.37 / 10.6 / 18.0	7.70 / 12.1 / 21.0
		Quality of synthetic data			
	Synthetic pair	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]
	clb (synthetic)	6.40 / 3.3 / 6.0	6.44 / 3.5 / 6.4	6.67 / 3.1 / 5.7	7.31 / 4.3 / 6.3
	slt (synthetic)	6.32 / 4.1 / 6.8	6.32 / 3.7 / 6.5	6.40 / 4.5 / 7.7	7.01 / 8.6 / 13.1
	Mean of MCD	**6.36**	**6.38**	**6.54**	**7.16**
	bdl (synthetic)	6.88 / 5.1 / 8.5	6.78 / 5.0 / 8.4	6.96 / 4.1 / 6.4	7.68 / 4.8 / 7.4
	rms (synthetic)	6.61 / 4.0 / 8.0	6.79 / 4.1 / 7.3	7.05 / 4.4 / 8.1	7.49 / 6.9 / 11.0
	Mean of MCD	**6.75**	**6.79**	**7.00**	**7.59**

5. <source synthetic and source natural, target natural and target synthetic> ((synthetic + natural)–(natural + synthetic))

Table A.1 lists the results from different sizes of training data. Female-speaker pair (clb, slt) and male-speaker pair (bdl, rms) were used for VC training. Here, mean of MCD represents the quality of synthetic data which also reflects the performance of corresponding TTS models. The overall quality of synthetic data produced by female speakers was better than that produced by male speakers in each datasize. This is because the original TTS model which was pretrianed by the data of the female speaker (judy) in the M-AILABS database [183], inherits the common characteristics of female speakers and better adapted to synthesizing female speech.

The experimental results show that the VC results from using pure natural parallel data were always the best among the five groups. On the contrary, the VC results

were the worst by only using SPD for training. From the overall trend, it is obvious that larger datasize lead to better TTS performance.

In general, under each training datasize, synthetic–natural tended to show the best VC results among the three training pairs using SPD. However, the difference of the VC results between natural–synthetic and synthetic–natural tended to be marginal as the datasize increased. As the datasize decreased, the gap between the results of synthetic–natural and natural–synthetic gradually increased. These results suggest that the source side is more robust against the quality of the synthetic data than the target side. Therefore, caution should be taken on the quality of the synthetic data to reduce the negative impact when the performance of TTS drops.

In addition, the VC results of mixed training pairs ((synthetic + natural)–(natural + synthetic)) became the second best when the datasize of female-speaker pair (clb–slt) was greater than or equal to 400. The reason is that the synthesized speech is good enough. Hence, the mixed training pairs have a larger pool of high-quality data. On the other hand, we can see from the male-speaker pair (bdl–rms) that when the datasize was 450, the VC result of the mixed training pair was still slightly worse than that of synthetic–natural, which is different from the result obtained from the female-speaker pair with the same datasize, indicating that the performance of the TTS models was not sufficiently high.

We can conclude that in general, the TTS performance is most critical in terms of the impact on the VC results. For Q1-1, it can be clarified that when the original datasize is small, the quality of SPD will be degraded due to the limited performance of the TTS model. Especially when the target speaker used SPD or the entire training set only contained SPD, the VC result was in general, unsatisfactory. Conversely, we can appropriately reduce the constraints on the use of SPD when the SPD quality is good enough, and use it together with natural data to ensure that the VC training datasize is large enough to achieve better VC results. For Q1-2, the training dataset of synthetic–natural, in general, performed better. However, when the TTS performance

was good enough, the mixed training pairs brought the optimal results.

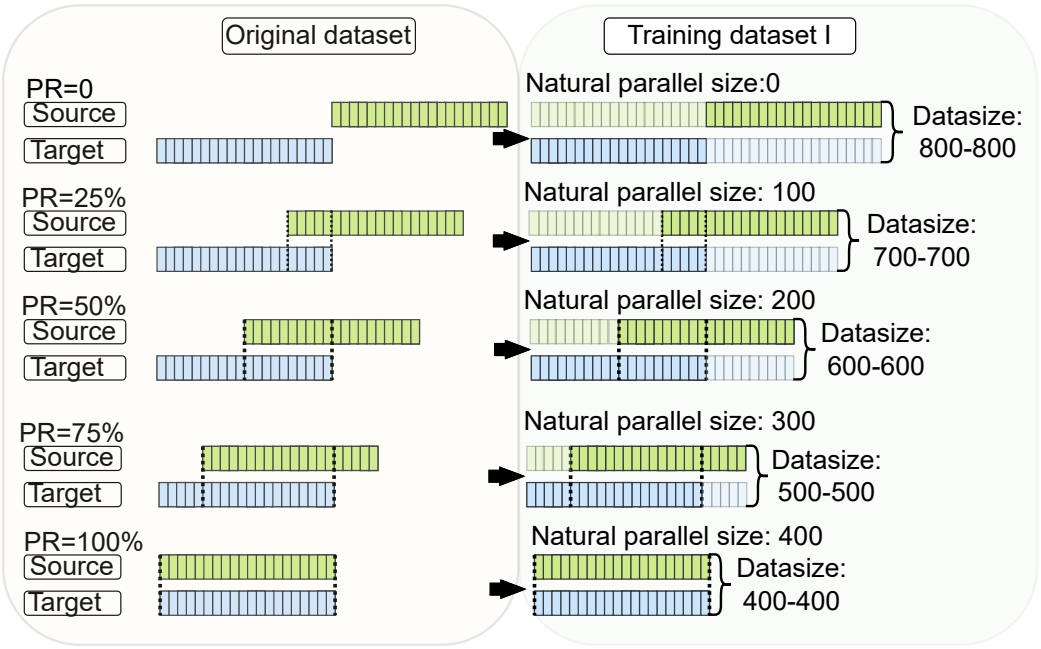
A.1.3 Investigation for Semi-parallel Setting

This part contains the investigation and experimental results from the semi-parallel setting based on Q2. Parallel Ratio (PR) is used to represent the proportion of the natural parallel corpus, so as to reflect the semi-parallel setting. The PR of training data pairs were gradually increased from totally non-parallel (PR = 0%) to parallel (PR = 100%) with the same datasize of the original training data. The respective TTS models of source speaker and target speaker were trained in case of constant datasize but a different semi-parallel setting for each group, which means the sum of datasize of VC parallel training involving synthetic data and natural data is different. Different sizes of original training data were selected (datasize = 400, 40) to compare the VC results. Figure A.2 illustrates the procedure with the original datasize of 400. The experiment is divided into two parts: Using all feasible data for training, as shown in Figure A.2(a), and further removing the natural–synthetic part of the semi-parallel case for training, as shown in Figure A.2(b).

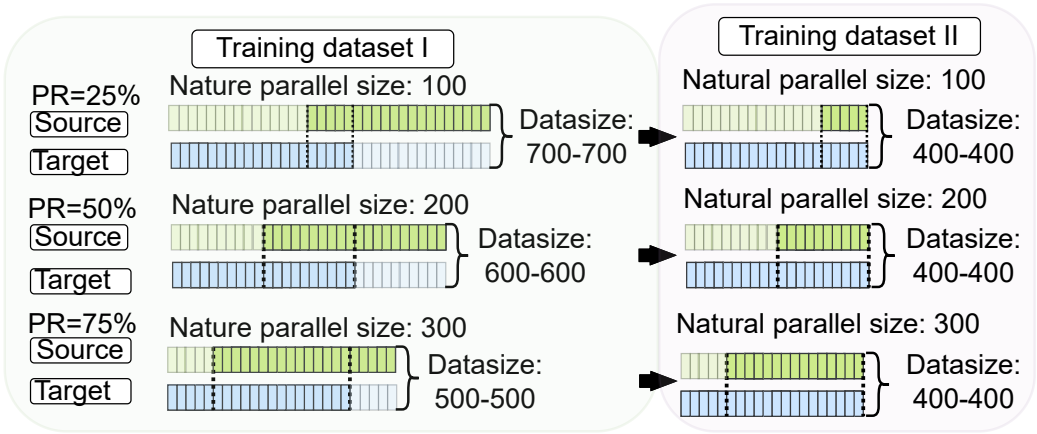
The results of the experiments are listed in Table A.2. Without cutting training data, we can see that as PR changed from 0% to 100%, the results became better when the original datasize was 40. In addition, except for the case of PR = 1, the results with the original datasize of 400 were relatively better at PR = 0, but the overall trend did not change significantly. When the natural–synthetic part was removed, the semi-parallel training results with the original datasize of 40 was improved, which is contrary to the original datasize of 400.

Based on the series of experiments, we can conclude that the outcomes are related with the original training datasize as:

1. When training-data level of TTS model was 400, the quality of synthetic data generated from TTS models was high. If PR = 0, the amount of VC training data was 800–800. Hence, the results were improved by providing large data-



(a) Training dataset I retains all SPD and the original data for training.



(b) Training dataset II removes natural–synthetic part.

Figure A.2: Training procedure with different Parallel Ratio (PR)-based semi-parallel settings (e.g., datasize=400).

Table A.2: Experimental results under different semi-parallel setting.

Original datasize: 400					
Speaker Source–Target	Parallel ratio [%]	Training size	MCD / CER / WER [dB] / [%] / [%]	Eliminated training size	MCD / CER / WER [dB] / [%] / [%]
clb–slt	0	800–800	6.59 / 3.7 / 7.1	—	—
	25	700–700	6.61 / 4.5 / 9.2	400–400	6.85 / 2.9 / 5.0
	50	600–600	6.67 / 8.0 / 13.8	400–400	6.73 / 3.4 / 5.8
	75	500–500	6.64 / 2.4 / 4.9	400–400	6.74 / 3.3 / 6.4
	100	400–400	6.35 / 5.0 / 9.1	—	—
Original datasize: 40					
Speaker Source–Target	Parallel ratio [%]	Training size	MCD / CER / WER [dB] / [%] / [%]	Eliminated training size	MCD / CER / WER [dB] / [%] / [%]
clb–slt	0	80–80	7.69 / 12.2 / 20.6	—	—
	25	70–70	7.41 / 8.5 / 15.8	40–40	6.94 / 6.7 / 12.2
	50	60–60	7.07 / 5.7 / 10.1	40–40	6.81 / 6.0 / 12.6
	75	50–50	6.86 / 4.0 / 8.4	40–40	6.80 / 5.5 / 10.5
	100	40–40	6.80 / 4.1 / 8.9	—	—

size and high quality of synthetic data. With the increase of PR, the increase of the natural–natural part and the decrease of the synthetic datasize can compensate for the negative effect caused by the decrease of the total training datasize. However, when the TTS performance is good enough, the datasize will become the critical factor. This also explains why further cancellation of the natural–synthetic part reduced the VC performance.

2. As the training datasize was small (e.g., 40), the resulting trained TTS models were unable to generate high-quality data. Therefore, the conversion results should be poor when PR is low. Here, adjusting the training datasize while keeping the usage of synthetic data is unable to compensate for the negative impact from the bad TTS performance. If PR and the ratio of using natural–natural were higher, the results would become better. Even so, the best results should not be better than any experimental result in which the datasize is 400, since the total amount of data is much less. Meanwhile, we can see that in the case of the semi-parallel setting, eliminating the natural–synthetic part improved the VC results, and with the increase of PR, the performance was close to natural–natural

Table A.3: Experimental results of adding external data with different datasizes. TTS-400 and TTS-200 represent homologous datasize of TTS finetuning.

Speaker	**External synthetic speech quality**	**TTS-400**			
clb–slt	Source / Target	Natural–Natural	Natural–Synthetic	Synthetic–Natural	Synthetic–Synthetic
External Datasize	WER / WER [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]
–	–	**6.35 / 5.0 / 9.1**	**6.64 / 3.7 / 7.5**	**6.68 / 3.4 / 6.7**	**6.85 / 4.8 / 8.5**
1k	0.0 / 0.0	6.19 / 1.8 / 4.6	6.48 / 3.0 / 6.6	7.18 / 8.4 / 12.1	8.13 / 27.3 / 31.4
2k	0.0 / 0.0	6.15 / 1.6 / 4.0	6.52 / 3.0 / 7.1	6.69 / 3.9 / 6.6	8.29 / 32.7 / 36.8
5k	0.9 / 0.9	**6.22 / 1.8 / 4.4**	**6.50 / 2.9 / 6.3**	**7.12 / 8.4 / 11.8**	**7.35 / 12.3 / 15.5**
Speaker	**External synthetic speech quality**	**TTS-200**			
clb–slt	Source / Target	Natural–Natural	Natural–Synthetic	Synthetic–Natural	Synthetic–Synthetic
External Datasize	WER / WER [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]	MCD / CER / WER [dB] / [%] / [%]
–	–	**6.66 / 5.3 / 9.5**	**6.74 / 5.6 / 10.7**	**6.68 / 3.1 / 6.5**	**6.97 / 8.5 / 13.1**
1k	0.0 / 0.0	6.46 / 2.2 / 4.9	6.66 / 4.0 / 7.4	6.69 / 3.9 / 7.9	7.72 / 17.9 / 22.3
2k	0.0 / 0.0	6.50 / 2.1 / 5.0	6.66 / 3.3 / 7.4	6.83 / 3.4 / 6.4	7.67 / 18.3 / 21.8
5k	0.6 / 0.8	**6.40 / 3.0 / 6.3**	**6.65 / 3.9 / 8.5**	**7.01 / 6.1 / 10.7**	**7.78 / 19.6 / 24.8**

(PR = 100%) training. So the natural–synthetic part had a great negative impact on VC when the datasize was small, which is consistent with the conclusion of Q1.

A.1.4 Investigation on the External Text Data

This subsection provides an experimental study to verify the influence of external text data corresponding to Q3. The original non-parallel dataset was fixed to train the TTS models. Then all the external text data from the M-AILABS database (15,369 utterances in total) [183] was input to TTS models to generate the external SPD. WER was used as the basis for selecting the highest quality SPD. The external SPD with different sizes were input into four non-parallel training cases.

Table A.3 presents the results of different original datasizes. By comparing the VC results of non-external data, we can see that the introduction of external data had a positive impact on VC. Especially when the original training pair was natural–natural, introducing external data significantly improved the VC results. On the other hand, the natural–synthetic pair also showed better VC results after adding the external syn-

thetic data. With more external data involved in the training, the result tended to be better. Nevertheless, the influence of external data on the synthetic–natural pair showed an opposite trend to the former. As a result, we can conclude that natural–synthetic outperformed the synthetic–natural after adding the external data.

Finally, the addition of external data was detrimental when the original data was synthetic–synthetic. In other words, synthetic data incompletely inherits all the features of natural speech. On the contrary, natural data plays a corrective role. When natural part is lost, the VC model will fully learn the features of synthetic data, and external data enhances this learning process, leading to a worse effect.

Therefore, the composition of the original training dataset should be considered before introducing the external data. The original training datasets which are natural–natural or natural–synthetic can benefit from the external data. On the contrary, it is unnecessary to introduce external SPD under the circumstance that the original training pair contains source synthetic data.

A.1.5 Subjective Evaluation

This subsection takes samples generated by experiments of the objective evaluation to construct the test sets for subjective evaluation tests.

For the subjective tests of Q1, the synthetic speech of male and female target speakers (slt, rms) was fed into the test set. According to the naturalness and similarity results presented in Table A.4, using the synthetic–natural dataset was slightly better than that of using natural–synthetic under the datasize of 450. This difference of the performance became significant when the datasize was 80. Meanwhile, the performance of the method by using mixed training pair (synthetic + natural)–(natural + synthetic) was comparable with using synthetic–natural in the datasize of 450, but slightly worse in the datasize of 80.

For the subjective tests on Q2, the naturalness and similarity results are shown in Table A.5. The evaluation of the performance approached PR = 1 by increasing

Table A.4: Results of subjective evaluation using test set under 450 and 80 datasize with 95% confidence intervals for Q1. "n–n", "n–s", "s–n", "s–s", and "(s + n)–(n + s)" denote natural–natural, natural–synthetic, synthetic–natural, synthetic–synthetic, and (synthetic + natural)–(natural + synthetic), respectively.

Speaker	Training data	TTS-450		TTS-80	
	Description	Naturalness	Similarity	Naturalness	Similarity
clb–slt bdl–rms	n–n	3.67 ± 0.14	71% ± 8%	3.42 ± 0.22	61% ± 8%
	n–s	3.43 ± 0.15	57% ± 8%	2.91 ± 0.21	46% ± 7%
	s–n	3.52 ± 0.14	62% ± 8%	3.26 ± 0.20	53% ± 9%
	s–s	3.33 ± 0.16	52% ± 7%	2.81 ± 0.21	43% ± 9%
	(s + n)–(n + s)	3.55 ± 0.14	63% ± 8%	3.12 ± 0.23	49% ± 8%

Table A.5: Results of subjective evaluation using test sets with 95% confidence intervals for Q2.

Speaker	Training data	TTS-40	
	Description	Naturalness	Similarity
clb–slt	PR = 0%	1.95 ± 0.17	20% ± 7%
	PR = 50%	2.74 ± 0.19	47% ± 8%
	PR = 50% without natural–synthetic	3.68 ± 0.14	68% ± 7%
	PR = 100%	3.91 ± 0.12	78% ± 8%

Table A.6: Results of subjective evaluation using test sets under 400 and 200 datasize with 95% confidence intervals for Q3. "n–n", "n–s", "s–n", and "s–s" denote natural–natural, natural–synthetic, synthetic–natural, and synthetic–synthetic, respectively.

Speaker	Training data	TTS-400		TTS-200	
clb–slt	Description	Naturalness	Similarity	Naturalness	Similarity
Non-external data	n–n	3.71 ± 0.11	73% ± 7%	3.69 ± 0.11	65% ± 6%
	n–s	3.45 ± 0.11	61% ± 7%	3.43 ± 0.15	53% ± 8%
	s–n	3.67 ± 0.12	69% ± 8%	3.54 ± 0.15	60% ± 7%
	s–s	3.32 ± 0.15	58% ± 7%	3.34 ± 0.14	51% ± 7%
Adding-external data	n–n	4.10 ± 0.12	83% ± 6%	3.88 ± 0.13	69% ± 7%
	n–s	3.73 ± 0.14	70% ± 8%	3.55 ± 0.13	60% ± 7%
	s–n	3.57 ± 0.15	63% ± 8%	3.38 ± 0.12	51% ± 8%
	s–s	3.10 ± 0.18	52% ± 9%	2.80 ± 0.17	40% ± 8%

PR and removing the natural–synthetic part of the dataset simultaneously under the semi-parallel setting with small datasize.

For the subjective tests on Q3, the results of the subjective test with the datasize of 400 and 200 showed a synchronous trend in Table A.6. We can see that external SPD

improved the performance of natural–synthetic and natural–natural.

The overall results are consistent with the findings in the objective evaluations. The answers to the three questions, Q1, Q2, and Q3 mentioned at the beginning are summarized as follows:

- **A1:** SPD is feasible for seq2seq non-parallel VC. SPD is produced by the TTS models trained with the original dataset of source and target speakers. Therefore, the VC results using SPD are determined by the performance of TTS models and VC training datasize. When the original data are sufficient, we can obtain TTS models with excellent performance, resulting in a better VC result. In addition, the VC result is also affected by the object of using SPD. When the dataset is limited, providing SPD for the source speaker and retaining parallel natural data for the target speaker can yield a better VC result.

- **A2:** When the dataset is semi-parallel, we should try to ensure the PR is large enough. Under this premise, when the original datasize is large, the introduction of SPD into target speaker or source speaker can both achieve ideal VC results. Thus, the full use of all types of SPD to ensure the amount of data, can maximize the benefits. On the contrary, when the original datasize is small, well-performing TTS models are difficult to obtain. Introducing training pair with a negative impact such as natural–synthetic should be avoided.

- **A3:** The introduction of external text data can provide a large amount of useful parallel data for VC. External data can significantly improve the VC results when the original dataset is natural–natural or natural–synthetic. However, it should be noted that when there is no natural speech or only source natural speech in the original dataset, the introduction of external data will lead to a negative impact on VC training.

A.2 Summary

This supplementary study carried out a series of experiments to study the impact of using SPD on non-parallel seq2seq VC and to address three questions. The experimental results provide guidance for using synthetic speech. The results showed that SPD is feasible in the absence of natural parallel data, and the VC results are related to both the TTS performance and the VC training datasize. When the original datasize was larger, the effect of using SPD was better. In most cases, the VC results would benefit more by exclusively providing synthetic data to the source speaker than to the target speaker. However, this situation was reversed when the external data were added. Moreover, although the research systematically explored and verified the different cases of SPD on seq2seq VC, the main focus was on a limited number of speaker pairs. In addition, the maximum size of the original data was only 450, which means that we were unable to determine the upper limit of the training effect that SPD could achieve for the time being. Therefore, using more speakers and a larger amount of data to investigate the beneficial trend that seq2seq non-parallel VC could obtain from SPD is the future research direction. In terms of methodology, VC models which can directly process non-parallel data can be introduced to compare the performance with the way of using SPD on seq2seq VC in the future research, so as to further clarify the role of SPD.

Publications

Journal Papers

1. D. Ma, L.P. Violeta, K. Kobayashi, and T. Toda, "Pretraining and fine-tuning techniques for electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 33, pp. 3189–3201, July 2025.

2. D. Ma, Y. Choi, T. Fujimura, F. Li, C. Xie, K. Kobayashi, and T. Toda, "Sequence-to-sequence voice conversion-based techniques for electrolaryngeal speech enhancement in noisy and reverberant conditions," *APSIPA Transactions on Signal and Information Processing*, Vol. 14, No. 1, pp. e8_1–40, May 2025.

3. L.P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, and T. Toda. "Resolving domain mismatches in electrolaryngeal speech enhancement with linguistic intermediates," *IEEE Journal of Selected Topics in signal Processing*, Vol. 19, No. 5, pp. 827–839, June 2025.

4. F. Li, F. Shen, D. Ma, J. Zhou, L. Wang, F. Fan, T. Liu, X. Chen, T. Toda, and H. Niu, "Mandarin speech reconstruction from surface electromyography based on generative adversarial networks," Medicine in Novel Technology and Devices, Vol. 26, No. 100359, pp. 1–7, Mar. 2025.

5. F. Li, F. Shen, D. Ma, J. Zhou, S. Zhang, L. Wang, F. Fan, T. Liu, X. Chen, T. Toda, and H. Niu, "End-to-end Mandarin speech reconstruction based on ultrasound tongue

images using deep learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 33, pp. 140–149, Dec. 2024.

6. L. P. Violeta, D. Ma, W. C. Huang, and T. Toda, "Pretraining and adaptation techniques for electrolaryngeal speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 32, pp. 2777–2789, May 2024.

International Conferences

1. D. Ma, J. Mi, F. Li, L. P. Violeta, K. Kobayashi, and T. Toda, "Improving electrolaryngeal speech enhancement via a representation learning method based on integrated text and speech representations," in *Proceedings of 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC)*, 6 pages, Copenhagen, Denmark, July 2025.

2. D. Ma, Y. Choi, F. Li, C. Xie, K. Kobayashi, and T. Toda, "Robust sequence-to-sequence voice conversion for electrolaryngeal speech enhancement in noisy and reverberant conditions," in *Proceedings of 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC)*, 4 pages, Orlando, FL, USA, July 2024.

3. D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Two-stage training method for Japanese electrolaryngeal speech enhancement based on sequence-to-sequence voice conversion," in *Proceedings of IEEE Spoken Language Technology Workshop (SLT) 2023*, pp. 949–954, Doha, Qatar, Jan. 2023.

4. 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 *Proceedings of 13th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 870–877, Tokyo, Japan, Dec. 2021.

5. J. Mi, X. Shi, D. Ma, J. He, T. Fujimura, and T. Toda, "Two-stage framework for robust speech emotion recognition using target speaker extraction in human speech noise conditions," in *Proceedings of 16th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 6 pages, Macau, China, Dec. 2024.

6. F. Li, F. Shen, D. Ma, S. Zhang, J. Zhou, L. Wang, F. Fan, T. Liu, X. Chen, T. Toda, and H. Niu, "Mandarin speech reconstruction from tongue motion ultrasound images based on generative adversarial networks," in *Proceedings of 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC)*, 4 pages, Orlando, FL, USA, July, 2024.

7. L. P. Violeta, W. C. Huang, D. Ma, R. Yamamoto, and T. Toda, "Electrolaryngeal speech intelligibility enhancement through robust linguistic encoders," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2024*, pp. 10961–10965, Seoul, Korea, Apr. 2024.

8. L. P. Violeta, D. Ma, W. C. Huang, and T. Toda, "Intermediate finetuning using imperfect synthetic speech for improving electrolaryngeal speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023*, 5 pages, Rhodes, Greece, June 2023.

Technical Reports

1. D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Sequence-to-sequence voice conversion for electrolaryngeal speech enhancement with multi-stage pretraining and fine-tuning techniques," *IEICE Technical Report*, SP2023-32, pp. 27–32, Oct. 2023.

2. L.P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, T. Toda, "Electrolaryngeal speech enhancement through strong linguistic encoding methods," *IEICE Technical Report*, SP2023-33, pp. 33–38, Oct. 2023.

Domestic Conferences

1. D. Ma, L. P. Violeta, K. Kobayashi, and T. Toda, "Sequence-to-sequence voice conversion training using synthetic parallel data for electrolaryngeal speech enhancement," *2022 Autumn Meeting of Acoustical Society of Japan (ASJ)*, 2-8-8, Sep. 2022.

Others

1. F. Li, F. Shen, D. Ma, J. Zhou, L. Wang, F. Fan, X. Chen, T. Toda, H. Niu, "Mandarin speech reconstruction from neck and facial surface electromyography," in *Proceedings of 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Research poster presentation, Copenhagen, Denmark, July 2025.

Awards

1. 3rd Place Award in EMBC 2025 Student Paper Competition, issued by IEEE Engineering in Medicine and Biology Society, July 2025.

2. EMBS Student Paper Competition Finalist, issued by IEEE Engineering in Medicine and Biology Society, May 2025.

3. Best Paper Award, issued by the 16th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Dec. 2021.