APSIPA Transactions on Signal and Information Processing, 2025, 14, e8

Original Paper

Sequence-to-sequence Voice Conversion-based Techniques for Electrolaryngeal Speech Enhancement in Noisy and Reverberant Conditions

Ding Ma[1*], Yeonjong Choi[1], Takuya Fujimura[1], Fengji Li[2], Chao Xie[1], Kazuhiro Kobayashi[1,3] and Tomoki Toda[4]

[1] *Graduate School of Informatics, Nagoya University, Nagoya, Japan.*
[2] *School of Biological Science and Medical Engineering, Beihang University, Beijing, China.*
[3] *TARVO, Inc., Nagoya, Japan.*
[4] *Information Technology Center, Nagoya University, Nagoya, Japan.*

ABSTRACT

Electrolaryngeal (EL) speech is artificial speech produced using an electrolarynx to aid laryngectomees in communicating without vocal fold vibrations. Compared with normal speech, EL speech lacks essential phonetic features and differs in temporal structure, resulting in poor naturalness, speaker identity, and intelligibility. Sequence-to-sequence (seq2seq) voice conversion (VC) emerges as a promising technique for overcoming the challenges in EL-speech-to-normal-speech conversion (EL2SP). Nonetheless, most VC studies for EL2SP focus on converting clean EL speech, overlooking real-world scenarios where EL speech is interfered with background noise and reverberation. To address this, we propose a novel seq2seq VC-based training method. In contrast to relying on extra augmentation modules to tackle interferences, our method requires only a single framework. First, we pretrained a normal-to-normal seq2seq VC model, adapted from a text-to-speech model.

*Corresponding author: Ding Ma, ding.ma@g.sp.m.is.nagoya-u.ac.jp

Received 17 December 2024; revised 26 February 2025; accepted 26 March 2025
ISSN 2048-7703; DOI 10.1561/116.20240094

Then, we employed a two-stage fine-tuning in a many-to-one style leveraging pseudo-noisy and reverberant EL speech data generated from limited clean data. We evaluated several system designs of our method. The intermediate representations of these systems were also analyzed to understand their role in filtering the interferences. Comparative experiments demonstrated that our method significantly outperforms EL2SP baselines, non-trivially handling both clean and noisy-reverberant EL speech, which sheds light on possible directions for improvement.

Keywords: electrolaryngeal speech, sequence-to-sequence voice conversion, real-world scenarios, noisy, reverberant

1 Introduction

A speech utterance spoken by a healthy speaker consists of two types of information: (1) clear linguistic content, and (2) rich paralinguistic information [79] such as prosody, emotion, and speaker identity, which are essential for human speech communication. However, individuals with speaking disabilities cannot completely convey the above information and thus face serious communication barriers. One typical group of such individuals is laryngectomees. Although they retain the ability of how to speak, they cannot produce speech due to the permanent loss of important sound source organs, including vocal folds, after undergoing surgery to remove the larynx to treat laryngeal cancer [69]. To communicate, laryngectomees rely on an electrolarynx to simulate vibrations of the vocal folds to produce artificial speech, termed electrolaryngeal (EL) speech [61, 18, 80]. Unfortunately, there is a big gap between EL speech and normal speech owing to the two significant issues. First, intense noises from the high-energy excitation signals of the electrolarynx cause poor quality and unsatisfactory intelligibility of EL speech. Second, since the mechanically generated excitation signals cannot simulate variable F0 contours of healthy voice [44], EL speech sounds unnatural and robotic without any natural prosody or emotion. These limitations not only make EL speech difficult to understand but also cause discomfort for both users and listeners during communication.

Voice conversion (VC), which refers to a methodology that was originally designed to convert a speech from one speaker to another while preserving the underlying linguistic contents [9, 70, 49], has been applied as a promising enhancement approach for EL-speech-to-normal-speech conversion (EL2SP). Widely studied approaches to EL2SP focus on developing conventional statistic VC models that employ a *frame-wise* paradigm. Such models directly map

and convert the features of source EL speech to those of target normal speech frame-by-frame [52, 14, 68, 34, 35, 57], as depicted at the top of Figure 1. However, the fundamental problem of such a paradigm is that it relies on the explicit alignment of corresponding frames between source and target speech, forcing the temporal structure of converted speech to be the same as that of EL speech. It leads to the inaccurate inference of time-variant characteristics, such as rhythm and duration.

The sequence-to-sequence (seq2seq) model [64], which has emerged in recent years, provides a different strategy by decomposing the conversion paradigm [30], i.e., an encoder first disentangles linguistic contents from source speech features as intermediate representations. Through an attention mechanism, a decoder subsequently consumes these intermediate representations alongside target characteristics to reconstruct the features of converted speech, as shown in the middle of Figure 1. Thanks to such an encoder–decoder framework with the attention mechanism, seq2seq VC can automatically determine the duration of output and capture long-term dependencies such as prosody, suprasegmental characteristics of F0, and speaker identity [29]. Tanaka *et al.* [67] demonstrated that seq2seq VC surpasses conventional VC in normal-to-normal VC tasks. Coincidentally, recent advances in broader speaking-aid fields, such as dysarthric VC [25, 28], EL speech recognition [77, 76], and EL2SP [74, 87], are also largely attributed to the adoption of seq2seq models. Notably, the attempts described in [87] likewise highlight the advantages of the seq2seq model over non-seq2seq approaches for EL2SP. In summary, seq2seq VC shows promise in effectively bridging the complex alignment inherent in EL2SP.

Although the promising properties of seq2seq VC for EL2SP are exciting, there are two critical challenges that need to be addressed.

- **C-1: Constraints of limited data**. Most seq2seq VC models require a large amount of high-quality, parallel training data to ensure an ideal and generalized performance, whereas the available parallel data for EL2SP is practically low-resource. This results in the quality degradation of the converted speech, including mispronunciations and repeated/skipped phonemes [88].
- **C-2: Complexity of real-world scenarios**. Primary works on EL2SP depend on high-quality recordings from simple, clean environments to allow the model to focus on addressing the difficult mapping between EL and normal speech. However, in real-world scenarios, speech signals are often entangled with background interferences including noise and reverberation. The complex nature of the corruption of speech signals caused by interferences would make the distributions of speech-related information, such as linguistic contents and speaker identity, highly different from those in clean speech [78, 63]. On this basis, the current

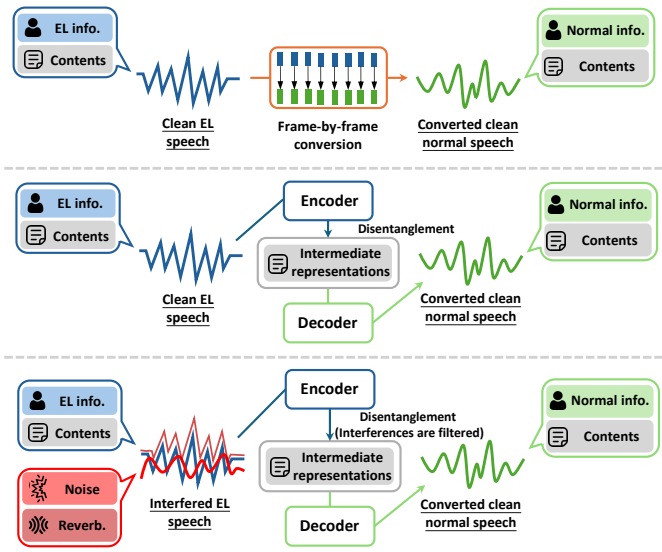


Figure 1: Overview of VC techniques for EL2SP. Top: conventional VC based on frame-wise function; middle: seq2seq VC that disentangles the spoken content of clean EL speech; bottom: seq2seq VC adapted to real-world EL2SP can generate intermediate representations by filtering out the interferences contained in EL speech.

approach would limit the models' capability to adapt to more complex real-world conditions.

A prevalent methodology for addressing C-1 involves transferring knowledge from a more accessible, large-scale dataset to a downstream fine-tuning dataset via a *pretraining–fine-tuning* scheme. As opposed to training from scratch, the compact, high-level representations from an extensive out-of-domain dataset can be leveraged to improve the performance of the target task, which has been widely validated in computer vision [58, 3], natural language processing [13], and speech signal processing [59].

Following this methodology and drawing inspiration from the Voice Transformer Network (VTN) [27], our prior work [43] developed a pretrained seq2seq VC model adapted from a text-to-speech (TTS) database of normal speech instead of acquiring a large VC corpus, which relaxes the high demand for parallel training data since only a single speaker's speech set is required. Furthermore, to reduce the restriction on transferability due to the huge domain shifts between EL2SP dataset and normal speech corpus, we employed a data augmentation approach to obtain more task-specific synthetic data (SD). Because the original EL2SP dataset is too small to build high-performance TTS models, we expanded it using low-quality parallel SD (PSD), in contrast to

most works [5, 38, 53] that depend on high-quality SD. This approach builds on the earlier findings [42], which demonstrated the viability of imperfect SD for normal-to-normal VC tasks. Our other focus [43] was to propose a novel two-stage fine-tuning technique to maximize useful knowledge from the extended EL2SP dataset while minimizing the accuracy drift due to PSD, thereby enhancing transfer learning efficacy.

For C-2, existing works are mainly from the perspective of extending the VC framework to improve its robustness for interfered conditions. A straightforward idea is to apply speech enhancement (SE) [71] as a preprocessing module. These works adopt an *enhancement–conversion* pipeline by incorporating extra components such as a denoising module (e.g., DCCRN [22]) and/or a dereverberation module (e.g., TasNet [40]), to develop noise- and/or reverberation-robust VC systems [83, 10]. However, such framework relies essentially on the prior knowledge of the interferences and the performance of the SE modules. Since the available SE modules are trained on normal speech, their effectiveness is compromised when applied to EL speech owing to the huge differences between EL and normal speech. Another critical drawback is that the speech information will be inevitably distorted during the processing and transmission of SE modules [85], adversely affecting the downstream VC. Moreover, most works employ frame-wise VC, which fails to solve the difficult alignments of EL2SP.

Considering the fundamental deficiencies of these sophisticated frameworks for real-world EL2SP, our resolution builds on the successes of previous efforts to overcome C-1. Starting from a new perspective of adaptation, we aim 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, we keep the seq2seq architecture and input EL speech 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 expect that non-speech interferences can be treated as extraneous information and filtered out, as depicted in the bottom of Figure 1. Another motivation is that the conversion target in our 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 we augment our training data 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 [43], 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.

In this study, we aim 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 our preliminary studies [43, 41], we not only design various systems in pursuit of state-of-the-art (SOTA) performance, but also present 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 this paper, we propose a many-to-one framework that integrates different noisy and reverberant EL SD with corresponding clean normal SD, enabling a simultaneous handling of multiple interferences without requiring any extra modules, labels, or strict alignment patterns.
- We design four systems of different fine-tuning levels on the basis of various EL data. Moreover, we follow the study outlined in [10], introducing both denoising and dereverberation SE modules to the EL2SP systems. Apart from using SE modules pretrained on normal datasets, we further fine-tune them on EL data through either cascading or joint training approaches, ensuring more optimized EL2SP baselines. Nonetheless, our SOTA system outperforms all these baselines.
- We extend the two-stage fine-tuning strategy, initially proposed for our seq2seq-based EL2SP systems, to SE network training. The experimental results confirm that this approach 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, we visualize the hidden representation spaces of learned EL2SP systems when dealing with real-world conditions and how those are related to the performance.
- We evaluate, through objective and subjective experiments, our systems under conditions of clean-, noisy-, and/or reverberant-EL speech. The reasonable results obtained in terms of speech quality, naturalness, and speaker similarity verify the generalizability of our methods.

2 Background and Related Works

2.1 Interference-robust VC

Following the discussion in Section 1, we hereafter detail three widely used categories for interference-robust VC: (1) statistical methods, (2) SE methods,

and (3) representation learning methods. Whereas the seminal works of latter two were both applied to TTS [71, 6],[21], we mainly focus on the most relevant papers on VC.

2.1.1 Interference-robust VC with Statistical Methods

Leveraging the sparse representations based on the non-negative matrix factorization (NMF) function [37] is a common statistical method for developing interference-robust VC. Takashima *et al.* [66] proposed an exemplar-based VC, wherein NMF decomposed 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, was 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.* [2, 1] endeavored to reduce the reliance on parallel data and improve training efficiency, their methods still cannot outperform other conventional VC models.

2.1.2 Interference-robust VC with SE Methods

SE methods, which involve connecting external SE modules or processing stages, are the mainstream approaches to address speech interferences. Miao *et al.* [48] realized a noise-robust VC but required 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-quefrency components. This method, although somewhat effective, relies heavily on intricate filtering techniques to handle noise throughout the VC process. In contrast, Xie *et al.* [83] 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 [72] 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.* [82] and Xie and Toda [81] 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 entailed 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.* [8], focusing on noisy condition, designed a lightweight SE module to achieve joint training with VC components incorporating generative adversarial networks (GANs) [17] 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 [10]. On the other hand, the authors of some VC studies [83, 10, 82, 81] affirmed that they did not rely on clean training data because of the use of the fixed SE module, but the SE pretraining still requires extensive clean data to ensure its performance.

2.1.3 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 [60, 20] provided some early inspirations in this direction. Afterwards, an attempt to address noisy condition using GAN-based domain adversarial training was proposed in one-shot VC [12], where two groups of gradient reversal layers and domain classifiers were assigned to the speaker and content encoders, respectively [15]. Therefore, training objective included not only the reconstruction loss but also domain classification losses. By learning encoded features that are invariant to noise, the framework is able to 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 studies considered overcoming the reverberation. Huang *et al.* [23] explored general interference-robust VC that combined adversarial and denoising training to tackle noise and reverberation. To achieve effective adversarial training, the embedding attack [24] was additionally used to generate adversarial samples, which were 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.* [51] suggested a VC framework that can overcome noisy-reverberant condition. However, besides acoustic signals, this framework requires transcriptions, which are consumed by extra phonetic and acoustic-automatic speech recognition (ASR) encoders for providing textual information, to enhance the efficiency of representation learning. In a similar study, Choi *et al.* [11]

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. However, most 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.

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

- Data processing. Since our 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. For clean EL2SP, Yang *et al.* [86] utilized synthesized EL speech, created by flattening the F0 contour of normal speech using the WORLD vocoder [50], to increase the volume of training data. Similarly, Xie *et al.* [83] and Huang *et al.* [24] developed interference-robust VC by simulating different interfered conditions from clean speech sets. However, these studies encountered a common limitation: Yang *et al.* [86] presumed the availability of ample EL speech for pretraining crucial components, while Xie *et al.* [83] and Huang *et al.* [24] relied on a large amount of high-quality clean data. In contrast, data augmentation in our method is a significant step towards more flexible and practical interference-robust EL2SP systems. We consider two types of data augmentation, i.e., increasing the amount of essential speech data according to imperfect PSD generated by fine-tuned TTS models, and then adding different real-world scenarios into the expanded EL data built.

- Framework. Our 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 recurrent neural networks [67] or convolutional neural networks [31], ours is structurally more suitable for handling real-world scenarios by leveraging the strengths of Transformer [73]. This not only accelerates training efficiency but also boosts a deep understanding of different types of data, as evidenced by its success across multiple datasets in large language models [7].

2.2 Transfer Learning in VC

Our method utilizes transfer learning to adapt to real-world scenarios. Hence, in this section, we provide an overview of the vast majority of methods used to improve transfer learning for VC techniques. Transfer learning requires an extensive pretraining dataset to achieve satisfactory generalization, but collecting a considerable amount of parallel VC corpus is difficult. Applying modules from TTS or ASR is a primary direction to attain effective pretraining, given the architectural commonalities with VC and the availability of corresponding large datasets. For instance, incorporating the attention and the decoder from pretrained TTS into VC has proven effective in generating high-fidelity speech [27, 89]. Similarly, transferring a pretrained encoder from ASR played a role in enhancing VC performance [26]. Building a more elaborate VC framework with extra components, such as pretrained phonetic posteriorgrams (PPGs) from ASR [90] and a text encoder of TTS [56], was also helpful for VC.

Our method neither requires an extensive, high-quality VC corpus nor relies on a complex structure. On the basis of our previous work [43], besides integrating a straightforward TTS pretraining, we designed a two-stage fine-tuning process combined with multiple types of interfered EL data, simulated using the imperfect clean EL SD.

3 Proposed Method

Recalling our main challenge—how to enhance the robustness of seq2seq EL2SP in the presence of interferences, especially when the original data is limited—the focus of our study is on combining easy-to-obtain SD (particularly simulated EL SD with varying acoustic properties) with transfer learning to improve the transfer performance in real-world environments. Figure 2 illustrates the process of developing our proposed method, which consists of two main parts: (1) Data augmentation (on the left) and (2) EL2SP training (on the right). Starting with the left, we fine-tune two TTS models for EL and normal speech using the original EL2SP dataset. We then use the two models to generate PSD, which contains EL and normal SD. Here, EL SD and the original EL data are further *interfered* by adding noise, reverberation, or a combination of both (in Section 3.1). Moving to the right, the pretraining–fine-tuning stages are carried out. We first develop a pretrained seq2seq VC model by TTS pretraining on a normal TTS database (in Section 3.2). 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 (in Section 3.3). Lastly, we construct various systems by adjusting the types of interfered data used

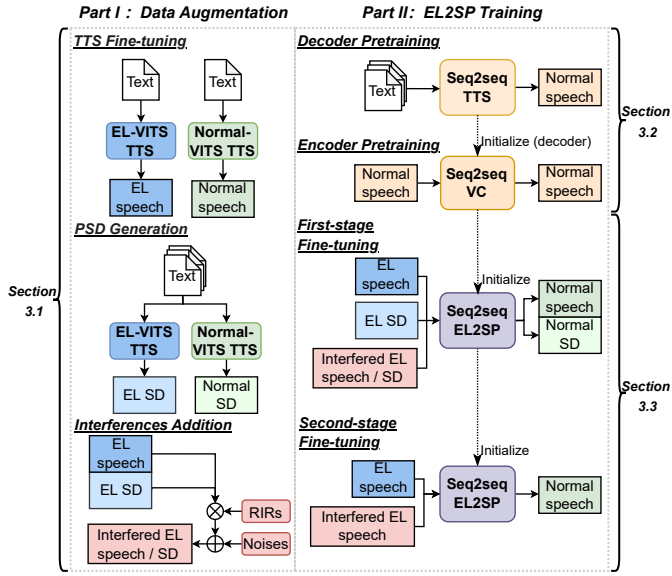


Figure 2: Overview of the proposed method for building real-world EL2SP, with the sections labeled to correspond to the description of each process.

during the fine-tuning process. Moreover, we design several typical baseline architectures by leveraging extra SE modules for fair comparisons with our method (in Section 3.4).

3.1 EL and Normal TTS Fine-tuning for PSD Generation

Given that only small-scale original data for EL2SP is available in this work, the process in this section aims to augment the training data by producing large-scale PSD. As the original dataset is too small to directly train a model, we fine-tune a pretrained VITS-based TTS model [33] via original source EL and target normal speech sets, yielding the corresponding EL-TTS and normal-TTS models, respectively. After this, we input the same external text set into both EL-TTS and normal-TTS models to generate PSD. Note that, owing to the low-resource dataset for TTS fine-tuning, the quality of PSD is poor.

Once the PSD is generated, we can obtain a much larger EL2SP dataset, 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.

3.2 Pretraining of Seq2seq VC Model

We adopt VTN [27] to build a pretrained *one-to-one* seq2seq VC model by utilizing Transformer-based TTS pretraining. Motivated by the same decoding mechanism between TTS and VC, we aim 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 process 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.

VTN pretraining includes 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 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.

3.3 Two-stage Many-to-one EL2SP Fine-tuning

At the beginning of the fine-tuning process, we employ the pretrained VTN to impart the effective valid *a priori* 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. We hence repeatedly use the original target normal data and normal SD to form the parallel pairs with their corresponding EL inputs. These training pairs are fed into the model to 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 essential 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 the above 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 discussed in Section 3.4.1.

3.4 Proposed Systems

3.4.1 Proposed System Conditions

As depicted in Table 1, we design four systems named *Model 1*, *Model 2*, *Model 3*, and *Model 4*. We emphasize that 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.

- *Model 1*: Since no SD is used, Model 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.
- *Model 2*: Model 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 Model 1 is applied.
- *Model 3*: Compared with Model 2, Model 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.
- *Model 4*: We argue that the types of EL data used in the second-stage fine-tuning may also impact the training performance. Therefore, in Model 4, on the basis of the consistent first-stage fine-tuning with Model 3, we conduct the second-stage fine-tuning that differs slightly by additionally leveraging the original EL inputs with only noise or reverberation.

3.4.2 Overview of the Composition of Comparable Baselines

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

Table 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
Model 1	***Yes/No***	C + NR	-	-
Model 2	***Yes/ Yes***	C + NR	***Yes/No***	C + NR
Model 3	***Yes/ Yes***	C + NR + N + R	***Yes/No***	C + NR
Model 4	***Yes/ Yes***	C + NR + N + R	***Yes/No***	C + NR + N + R

1) Baselines adapted to clean environment: As the first question, we look simply at how well our systems perform compared with those adapted solely to a clean environment. To contextualize this, we prepare two fundamental off-the-shelf baseline systems, named *Baseline 1* and *Baseline 2*: fine-tuning the pretrained VTN that is identical to our proposed systems, but without using any interfered data. Baseline 1 uses only the original clean pairs, whereas Baseline 2 additionally uses the same low-quality PSD as that of Models 2, 3, and 4.
2) Baselines using SE methods: Another question being considered is how effective our 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 proposed systems with Baselines 1 and 2, we also design four SE methods that are connected to the same EL2SP framework, to establish corresponding baseline systems. Note that these baselines, following the enhancement order demonstrated to achieve SOTA performance in [10], first conduct denoising and then dereverberation for noisy-reverberant inputs, as shown in Figure 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.

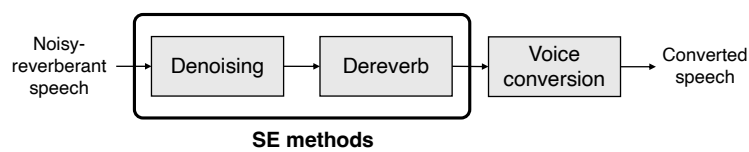
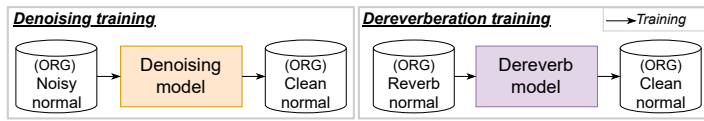
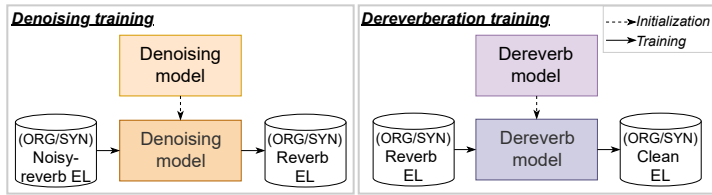


Figure 3: Overview of the baseline frameworks using SE methods.

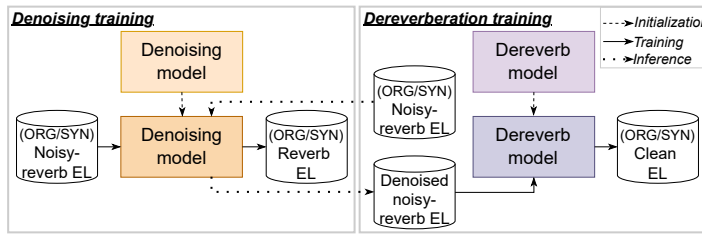
- *E-pt*: In Figure 4(a), E-pt provides a common SE strategy that involves using interfered data from widely available original normal human cor-



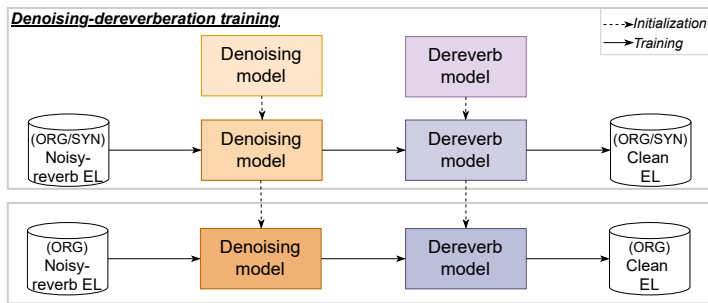
(a) Extension-pt (E-pt)



(b) Extension-ft (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: 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 is original or synthetic.

pora 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(b).

- *E-ft-c*: As plotted in Figure 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 we have indicated in Section 2, intermediate output would inevitably bring some distortions, potentially limiting SE performance. Therefore, E-ft-j, illustrated in upper block of Figure 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 our proposed seq2seq VC framework, we hypothesize that this approach can also be generalized to SE training. To demonstrate its generalizability, we further extend E-ft-j. Specifically, we adopt the same joint framework and initial training process as E-ft-j 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, we solely use the interfered and clean original data to refine the final SE model, thus establishing E-2ft-j, as illustrated in the lower block of Figure 4(d).

The SE methods are then individually ensembled with various EL2SP systems to deploy the corresponding baseline systems with higher comparative appeal. We specifically design three types of baseline systems based on above 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, we form four systems 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. We evaluate them 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, we specifically use E-ft-c and E-ft-j to separately process the interfered version of the training data from Baseline 1. The processed data is 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 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, we utilize E-ft-j and E-2ft-j as the SE modules and name the entire systems *E-ft-j-d-Base2* and *E-2ft-j-d-Base2*.

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

4 Experimental Evaluations

In this section, we first outlined the experimental protocol for our study (in Section 4.1), which comprises the datasets used, model architectures with their implementation configurations, and the metrics for evaluating the experimental results. Subsequently, we carried out 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 (in Section 4.2). The aspects we investigated include the effectiveness of our two-stage many-to-one fine-tuning, performance discrepancies under different conditions, and an analysis of the intermediate representations in our systems.

4.1 Experimental Protocol

4.1.1 Datasets

The three types of datasets used for the proposed method are as follows:

- *TTS database*: To accomplish the pretraining mentioned in Sections 3.1 and 3.2 for the seq2seq VC and VITS TTS models, we utilized the Japanese JSUT database containing 7696 utterances [62], which amounts to approximately 10 hours of speech. All transcriptions from the JSUT database were also selected for generating PSD.

- *Original EL2SP datasets*: To develop and evaluate EL2SP systems, we constructed two small-scale and semi-parallel EL2SP datasets, referred to as Patient 1 dataset and Patient 2 dataset, 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 Audacity recording software. All speakers involved in the recordings are native Japanese speakers.

  – *Patient 1 dataset* 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, a healthy male speaker recorded the normal speech under the same recording conditions.

  – *Patient 2 dataset* includes 573 EL utterances approximately 29 minutes, and 373 normal utterances 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, we leveraged fine-tuned TTS models to add the corresponding EL SD (213) for Patient 1 dataset and normal SD (200) for 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*: We leveraged 8109 and 8269 noise clips along with their corresponding room impulse responses (RIRs) from the WHAMR! dataset [45], to generate interfered EL data for Patient 1 and

Patient 2 datasets, respectively. We first created the reverberant versions of the original and synthetic EL data by convolving them with RIRs, whose T60 range was from 0.1 to 1.0 seconds. Subsequently, noise clips with five signal-to-noise ratios (SNRs) (0, 5, 10, 15, and 20 dB) were mixed 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, ensuring both the diversity of the dataset and the complete separation between training and test sets. We emphasize that all noise clips and RIR parameters in the training set, including those from original EL data and EL SD, were strictly different from those in the test set. Thus, the interferences in the test set were entirely unseen during model training.

Given the necessity for SE methods in establishing baselines, we additionally introduced two datasets for training the denoising and dereverberation models of E-pt. Following the work presented in [16], we used 1000 utterances from LibriSpeech [54], which were mixed with noise clips of the CHIME3 dataset [4], to conduct denoising training. Concurrently, still using the LibriSpeech dataset 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.

4.1.2 Configuration Settings

The EL data was 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 2048 FFT points and a 300-point shift to extract acoustic features. The implementations of the Transformer-based pretrained seq2seq VC and VITS TTS models were accomplished with the ESPnet toolkit [47, 19], following the official configurations. We additionally completed denoising training using a complex time-frequency mask (TFMask) network [32] and specifically referred to the configurations in [16]. Leveraging the Asteroid platform [55], the dereverberation module applied Conv-TasNet [39] as the backbone, following its original configurations. We utilized Parallel WaveGAN (PWG) neural vocoders [84] to reconstruct the waveforms of the EL2SP outputs, while the PSD 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, respectively.

4.1.3 Evaluation Metrics

1) Objective evaluation: We employed the following two objective evaluations for measuring different aspects of the evaluated EL2SP systems.

- MCD: The Mel cepstrum distortion (MCD, in dB) was used to measure the spectral distortions between the ground-truth target samples and the converted samples. This measurement, which can be viewed as an intrusive, L2-norm-based metric, is expressed as

$$\mathrm{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{D}\left(c_i^{(\mathrm{t})} - c_i^{(\mathrm{c})}\right)^2}, \tag{1}$$

  where D is the dimension of MCEPs, and $c_i^{(\mathrm{t})}$ and $c_i^{(\mathrm{c})}$ are the i-th dimensional coefficients of the target and converted MCEPs, respectively. A general assessment of quality performance can be verified on the basis of the MCD value, where a lower value indicates a higher quality of converted speech with less distortion.

- CER: We assessed the intelligibility accuracy and character consistency of converted samples using the character error rate (CER, in %). This metric, following a non-intrusive measurement, is calculated with an ASR engine trained as in [75].

In this study, we also explored the performance of the systems that use extra SE methods. Hence, we applied the scale-invariant signal-to-distortion ratio (SI-SDR, in dB) [36] and short-time objective intelligibility (STOI) [65] to evaluate the processed speech within the SE modules.

- SI-SDR: SI-SDR was 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 lower distortion and higher signal fidelity.

- STOI: STOI 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.

2) Subjective evaluation: Two subjective tests were carried out to evaluate the perceptual performance of the EL2SP systems on the generated speech.

- MOS test: An opinion test to assess naturalness was conducted in terms of the mean opinion score (MOS). During the test, listeners were asked to rate the naturalness of each given speech sample in a one-to-five scale (*5-Excellent, 4-Good, 3-Fair, 2-Poor, or 1-Bad*).
- SIM test: During the speaker similarity (SIM) test, listeners were presented with pairs of speech samples at the same time, one from normal target speech and one from test speech. Listeners were asked to judge whether they were spoken by the same speaker or not, by choosing one of four given levels: *Definitely the same*, *Maybe the same*, *Maybe different*, *Definitely different*.

Six randomly converted samples of clean and noisy-reverberant EL input from Models 1, 2, and 3, and Baseline 1 were chosen for each listener. Fifteen Japanese native speakers were recruited. Audio samples are available online[1].

4.2 Experimental Results and Analysis

4.2.1 Objective Evaluation Results

1) Comparison with seq2seq baselines: We compare all the proposed systems with the baselines trained on clean data, during which we simulate different acoustic properties on the EL test set to obtain conversion results under various interfered conditions. The results for Patient 1 dataset are documented in Tables 2, 3, and 4. Additionally, we conduct experiments on Patient 2 dataset, using the same set of representative interfered test conditions as in Table 2. The corresponding results are presented in Table 5. Models 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 Model 4. We caution that Models 3 and 4 share the same first-stage fine-tuning. Thus, for conciseness, the first-stage results of Model 4 are omitted from these tables.

We first examine the results for Patient 1 dataset. Significant advancements in the proposed systems over the baselines across all the interfered conditions are readily apparent in Tables 2, 3, and 4. In addition to this, when looking at Table 2, an interesting finding is noted under the metrics for converting clean EL speech. Here, owing to the use of low-quality PSD for training as well [43], Baseline 2 outperforms Model 1, with 6.15 versus (vs.) 6.44 in MCD, and 32.1 vs. 39.2 in CER. However, Models 2, 3, and 4 excel over Baseline 2, despite essentially leveraging the same volumes of speech data. Moreover, there is a consistent improvement observed from Models 1 to 4. We expect that incorporating a broader range of interferences into the EL SD will assist the model in learning more robust and discriminative features.

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

Table 2: Objective evaluation results based on 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 Models 2, 3, and 4, respectively.

Systems		**Clean EL**	**NR-EL**	**N-EL**	**R-EL**
		MCD / CER	MCD / CER	MCD / CER	MCD / CER
Model 1	–	6.44 / 39.2	7.11 / 55.9	7.08 / 54.3	6.82 / 42.3
Model 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
Model 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
Model 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 3: Objective evaluation results based on Patient 1 dataset, when the EL input is noisy-reverberant with the fixed SNRs (−5, 2, 12, and 22 dB).

Systems		**SNR: −5**	**SNR: 2**	**SNR: 12**	**SNR: 22**
		MCD / CER	MCD / CER	MCD / CER	MCD / CER
Model 1	–	8.70 / 73.2	7.59 / 63.7	6.82 / 46.5	6.63 / 47.1
Model 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
Model 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
Model 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

Table 4: Objective evaluation results based on 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 / CER	MCD / CER	MCD / CER	MCD / CER
Model 1	–	7.79 / 62.4	7.22 / 56.6	7.30 / 53.0	6.97 / 52.6
Model 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
Model 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
Model 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

These are beneficial for identifying linguistic information and thus enhancing the conversion of clean EL inputs. Conversely, Baseline 2 performs worse than Baseline 1 under the interfered input conditions. We argue that using low-quality PSD reduces the generalizability of the system to real-world scenarios if there are significantly large environmental mismatches.

We subsequently examine the conversion results of tests under interfered EL input conditions shown in Tables 2, 3, and 4. On this broader evaluation suite, the benefits of our proposed method are clearer, as consistent improvements can be observed from Models 1 to 4. Among the models shown in the

Table 5: Objective evaluation results based on 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 / CER	MCD / CER	MCD / CER	MCD / CER
Model 1	–	7.36 / 47.1	7.77 / 54.0	7.71 / 52.7	7.68 / 52.4
Model 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
Model 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
Model 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

tables, Model 4 shows the highest performance in 17 out of the total 22 results (covering MCD and CER metrics) across all eleven conditions, whereas Model 3 mainly takes 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 Models 2, 3, and 4, the second-stage fine-tuning mainly optimizes the results of the first stage, where MCD has a clearer optimization than CER (e.g., for "NR-EL" in Table 2, MCD / CER decrease from 6.33 / 38.4 to 6.09 / 36.7 for Model 2, and from 6.30 / 37.4 to 6.06 / 35.9 and 6.02 / 34.6 for Models 3 and 4, respectively). We are aware 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, we focus on evaluating the overall results of Models 2, 3, and 4 in the three tables. Generally, Model 3 outperforms Model 2 more often than not for each fine-tuning stage, whereas Model 4 advances further than Model 3. From the results in "N-EL" and "R-EL" categories in Table 2, the performances of these three reveal consistent improvements compared with those under the noisy-reverberant condition, and nearly match the performance under the clean EL condition. Tables 3 and 4 show a similar trend. Although Models 2, 3, and 4 show 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 are still significantly better than Model 1 and other baselines. Moreover, the conversion performances gradually improve and reach their best as the noise/reverberation intensity decreases, as indicated by the results at a higher SNR or lower T60 in Tables 3 and 4. The above findings indicate that our methods perform reasonably well under a wide range of interfered conditions and adapt particularly well to a single-interference condition or a noisy-reverberant condition with relatively mild

noise/reverberation effects. On the other hand, we notice that the overall results of Models 2, 3, and 4 are mainly better in the "R-NL" category than in the "N-EL" category in Table 2, suggesting that noise more detrimentally impacts EL2SP performance than reverberation. As the reverberation tends to stretch the length of EL speech, our models, 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 our methods to recognize varying intensities of reverberation and noise, thereby enhancing their robustness in real-world scenarios.

Furthermore, we examine the results for Patient 2 dataset in Table 5. Baseline 2 exhibits better MCDs for interfered conditions than Baseline 1, which differs from the observations in Patient 1 dataset. We attribute this to the larger size of the original EL data in Patient 2 dataset, 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 proposed systems align closely with those from Patient 1 dataset (Table 2): (1) Across all test conditions, the proposed systems significantly outperform the baseline systems and handle interfered conditions well. (2) Incorporating SD with two-stage fine-tuning leads to continuous improvements in system performance, ultimately yielding the best-performing model. (3) Comparing different proposed models, we find that during the two-stage fine-tuning, introducing augmented data with more fine-grained interference properties can further enhance model performance, making Model 4 the optimal system. By and large, these findings further verify the generalizability and robustness of our methodology across different EL2SP scenarios.

2) Comparison with seq2seq baselines using extra SE modules: Given that the noisy-reverberant condition represents the most severe challenge, we leverage Patient 1 dataset to provide a detailed summary and analysis of the performance differences among Baseline 1 and the baselines using SE modules under this condition, as shown in Table 6. Also, as shown in Table 7, we carry out experiments similar to those outlined in [10] 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 presents more negative results in Table 7. The quality of the processed speech continues 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

Table 6: Comparison results of baseline systems using SE training methods when the EL input from Patient 1 dataset is noisy-reverberant. Baseline 1 represents a lower bound.

Baseline systems	**NR-EL**
	MCD / 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

Table 7: Evaluation results of the modules based on different SE methods, according to the comparison between processed and clean EL data from 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 a lower bound.

SE modules	**Comparison**	**Evaluation metrics**	
		SI-SDR	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

leads to accumulated errors during enhancement processing. Conversely, E-ft and E-ft-c, both fine-tuned using additionally interfered EL SD, demonstrate improved performance compared with the lower bound, suggesting that even using imperfect SD also aids SE training. E-ft-j exhibits a further enhanced performance for both SI-SDR and STOI, which demonstrate the effectiveness of the joint training we proposed. Furthermore, E-2ft-j, which extends E-ft-j through two-stage fine-tuning, achieves the SOTA performance among all SE methods, with SI-SDR and STOI scores of 10.86 and 0.73, further verifying that the two-stage fine-tuning strategy is effective not only for VC but also for SE training.

In Table 6, the poor SE effect of E-pt also affects 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, exhibit a progressive optimization trend consistent with the performance of the SE methods documented in Table 7, reflecting their intrinsic reliance on the performance of SE modules. However, the improvement of these systems remains 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 models (see "NR-EL" results in Table 2), both unsurprisingly achieve better performance than Model 1 in terms of MCD and CER. Surprisingly, although E-ft-j-d-Base2 is further improved by additionally using processed EL SD with the same speech volume as Models 2, 3, and 4, it reaches the equivalent performance as the first-stage fine-tuning of Model 2 (6.34 vs. 6.33 in MCD, and 38.8 vs. 38.4 in CER), and still clearly underperforms the second-stage fine-tuning of Models 2, 3, and 4. Furthermore, thanks to the SOTA SE framework, E-2ft-j-d-Base2 further enhances performance, achieving MCD and CER of 6.25 and 36.9, respectively. However, it still falls short compared to the second-stage fine-tuning of Models 2, 3, and 4.

To further validate the effectiveness of joint-framework-based SE modules, we also used Patient 2 dataset to develop E-ft-j and E-2ft-j. Similarly, we compared the baselines extended with these SE modules, namely E-ft-j-d-Base2 and E-2ft-j-d-Base2. The experimental results are documented in Tables 8 and 9. The findings are aligned with those in Tables 7 and 6, reinforcing the effectiveness of the joint framework and the two-stage fine-tuning approach.

Table 8: Evaluation results of E-ft-j and E-2ft-j, according to the comparison between processed and clean EL data from Patient 2 dataset. The result between NR inputs and the corresponding clean speech is used as a lower bound.

SE modules	**Comparison**	**Evaluation metrics**	
		SI-SDR	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 9: Comparison results of baseline systems using SE training methods when the EL input from Patient 2 dataset is noisy-reverberant. Baseline 1 represents a lower bound.

Baseline systems	**NR-EL**
	MCD / 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

The overall comparative study makes the validity of our model clearer. We enhance the seq2seq VC framework 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, our model efficiently achieves SOTA performance, surpassing the widely used frameworks that rely on extra SE modules.

4.2.2 Subjective Evaluation Results

Figure 5 depicts the MOS and SIM results, incorporating eleven types of speech mixed into the respective test set based on Patient 1 dataset. Both metrics exhibit a smooth optimization trend in the tasks of converting clean and noisy-reverberant EL speech. Consistent with the results of objective evaluations, the proposed systems significantly outperform Baseline 1, with a progressive enhancement from Models 1 to 3. Particularly in noisy-reverberant EL2SP, Models 2 and 3 reveal the closer speaker similarity and naturalness to the target compared with Model 1, underscoring the effective robustness achieved through the two-stage fine-tuning with interfered SD. Moreover, the edge of Model 3 over Model 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 Models 2 and 3 yield results under the noisy-reverberant condition comparable in naturalness and speaker similarity to those under the clean condition. This showcases the superiority of our models in converting more natural speech with a closer speaker identity to the target under the noisy-reverberant condition.

4.2.3 Visualizations of the Hidden Representation Spaces

Since our technique is expected to assist the model's encoder in filtering out interferences and extracting speech-related knowledge, particularly linguistic representations, we provide visual evidence by conducting uniform manifold projections [46] to further demonstrate this. Leveraging Patient 1 dataset, we specifically analyze Models 1, 2, and 3, and use the results of Baseline 1 as a lower bound. Hidden representations of the test sets are extracted from the trained encoders of these systems. These representations are then visualized at utterance and phoneme levels, as shown in Figures 6 and 7, respectively. Note that we assess the encoding results under four conditions, namely, clean, noisy, reverberant, and noisy-reverberant.

1) Utterance-level visualization: Besides the impact of environmental interferences on encoding effects, the difference in linguistic content across utterances is the critical variable, affecting utterance-level representations. In this context, we find that Baseline 1 presents a relatively clear clustering effect for clean EL inputs, but the hidden representation space for interfered EL inputs exhibits poor discriminability. This suggests that Baseline 1 cannot adapt to the interfered scenarios.

Models 1, 2, and 3 show the roughly similar representation spaces, yet there are notable differences that warrant further analysis. Compared with Baseline 1, Model 1 shows better clustering. It is not difficult to infer that,

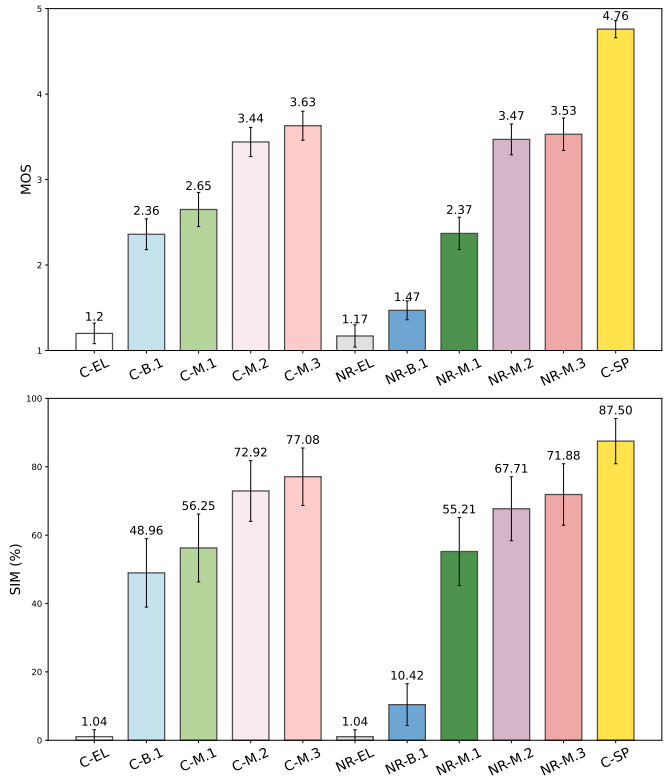
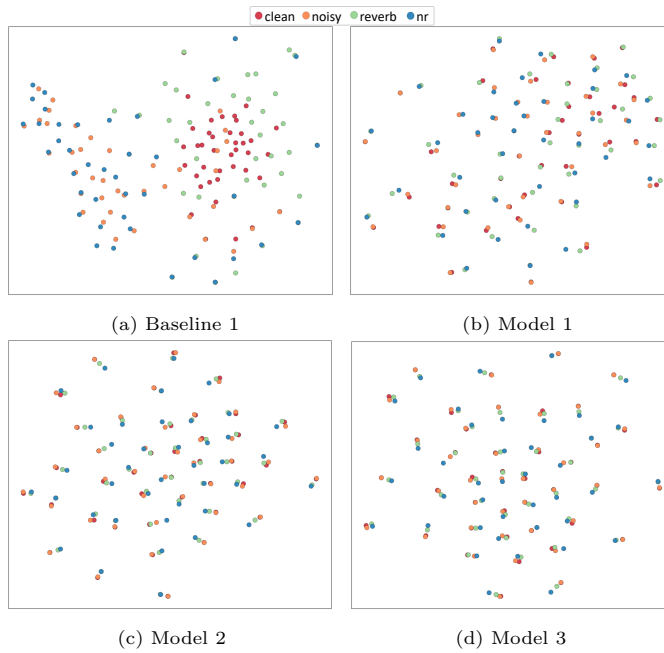


Figure 5: MOS (upper) and SIM (lower) results with 95% confidence interval under clean (C) and noisy-reverberant (NR) conditions, where B.1, M.1, M.2, and M.3 denote the outputs of Baseline 1, Model 1, and the second-stage outputs of Models 2 and 3, respectively. Clean, noisy-reverberant EL, and target normal speech are also used as the lower and upper bounds, denoted as C-EL, NR-EL, and C-SP, respectively.

as interferences are eliminated during encoding, speech with the identical linguistic content tends to cluster closely. Nevertheless, some areas of the hidden representation space still show weak separation performance, indicating that Model 1 struggles to effectively differentiate between features of different sentences owing to its performance limitations and the impact of interferences. Conversely, Models 2 and 3 exhibit more effective clustering. Both of them not only achieve more compact clustering for the same utterance contents but also form well-separated clusters for different utterance contents. These observations closely match our expectation, demonstrating the effectiveness and potential of our method based on the seq2seq architecture.



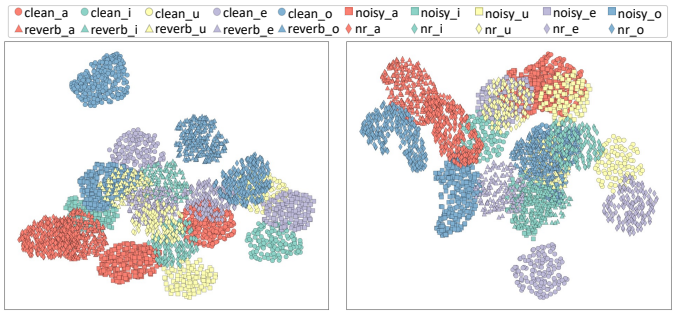
(a) Baseline 1 (b) Model 1

(c) Model 2 (d) Model 3

Figure 6: Visualizations of utterance-level hidden representations extracted from Baseline 1, Model 1, and the second-stage fine-tuning of Models 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.

2) Phoneme-level visualization: Since in this study, we utilize Japanese dataset, we color the five most common Japanese vowel phonemes and their corresponding hidden representations to simplify the plots. As shown in Figure 7, 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.

In Baseline 1, the five phoneme representations from the clean condition are relatively discretized. However, across all conditions, there is considerable overlap among different representations, suggesting that Baseline 1 does not distinctly differentiate phoneme features. Model 1 shows a clearer distribution of the same phonemes, yet overlaps still persist, albeit slightly improved from Baseline 1. Model 2 further exhibits 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.,



(a) Baseline 1 (b) Model 1

(c) Model 2 (d) Model 3

Figure 7: Visualizations of phoneme-level hidden representations extracted from Baseline 1, Model 1, and the second-stage fine-tuning of Models 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.

the phonemes "a" in reverberant speech and "e" in noisy-reverberant speech). Looking at Model 3, we find that the visualization of some phoneme representations, such as the phonemes "a" and "u", show a high degree of cluster purity, indicating the effective filtration of interferences and enhanced phoneme recognition. In addition, Model 3 ensures that the same phoneme types under different conditions cluster closely, and the overlaps between various phoneme representations are almost negligible. It demonstrates the clear delineation and robust grouping for the phonetic features of EL inputs across environmental variations. However, note that Models 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 our models 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 Models 1 to 3 underscores the efficacy of our method.

5 Discussion and Conclusion

In this study, we developed and evaluated the training strategy in a seq2seq framework to address two critical issues for the EL2SP task: (1) the practically low-resource EL2SP data available, and (2) the lack of adaptation to real-world interfered conditions. This work was based on interpreting encoding mechanism of seq2seq VC and the transfer learning. Aside from a pretraining process that leverages the attention and the speech decoder from a normal TTS model, we developed a unified, two-stage fine-tuning technique to address both problems simultaneously. During this process, we tackled the minimal-resource data by incorporating readily available, low-quality PSD. More crucially, by injecting fine-grained interferences into EL SD as the additional training materials to construct many-to-one mappings, we improved model's generalization 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 our training framework, we designed several systems at different levels. Experimental results demonstrate that our systems outperform the baseline systems trained on clean data in the EL2SP task under different test conditions. Furthermore, we systematically compared our method with mainstream methods using external SE modules. Although the performance of these baseline systems can be continuously improved by optimizing the SE architectures, the SOTA system in our approach still prevails in direct comparisons.

In addition to outperforming the baselines, another advantage of our method manifests in its simpler architecture. Adding new modules or architectures often increases complexity and separately handles interference elimination and conversion. Such approaches depend on the performance of these modules, whose networks require considerable training data. The above factors pose difficulties for actualization in practical scenarios. In contrast, our methods rely solely on a single seq2seq architecture. Moreover, the proposed fine-tuning method using SD is efficient and easily applicable in real-world environments.

We hope that our research provides a relatively fresh perspective for researchers, promoting cost-effective training approaches based on advanced model architectures that enhance adaptation to downstream tasks with limited resources, and the capability to disentangle non-essential knowledge. At present, the performance of our method under severe interferences still leaves room for much improvement compared with its performance under the clean

condition. In addition, continuing to investigate the potential of seq2seq techniques, optimizing our method, and expanding its applicability domain will be our other research directions in the future, which specifically contain the following aspects:

- We plan to enhance the generalizability of our method, including (1) Expanding our dataset to include more EL speakers to develop a multi-speaker EL2SP system. (2) Bridging our method to real-world applications by optimizing and deploying the proposed models with lower latency on mobile platforms, such as smartphones and iPads. This would facilitate more accessible speech conversion for laryngectomees.
- Integrating multimodal features to improve semantic integrity and speech quality is another promising direction. For instance, textual and visual modalities offer clearer linguistic and contextual cues than speech alone, potentially enhancing conversion quality in real-world environments. Therefore, we also plan to develop an multimodal EL2SP dataset.

6 Acknowledgements

This work was partly supported by JST CREST JPMJCR19A3 and AMED JP21dk0310114, Japan.

References

[1] 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*, 2015, 2015, 1–9.

[2] 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*, 97(6), 2014, 1411–8.

[3] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers", *arXiv preprint arXiv:2106.08254*, 2021.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME-speech separation and recognition challenge: Dataset, task and baselines", in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2015, 504–11.

[5] 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 *Interspeech*, 2019, 4115–9.

[6] 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 *Interspeech*, 2016, 352–6.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners", *Advances in Neural Information Processing Systems*, 33, 2020, 1877–901.

[8] 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.

[9] D. Childers, B. Yegnanarayana, and K. Wu, "Voice conversion: Factors responsible for quality", in *ICASSP 1985-1985 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 10, IEEE, 1985, 748–51.

[10] Y. Choi, C. Xie, and T. Toda, "An evaluation of three-stage voice conversion framework for noisy and reverberant conditions", in *Interspeech*, 2022, 4910–4.

[11] Y. Choi, C. Xie, and T. Toda, "Reverberation-controllable voice conversion using reverberation time estimator", in *Interspeech*, 2023, 2103–7.

[12] J.-C. Chou, C.-C. Yeh, and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization", in *Interspeech*, 2019, 664–8.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[14] 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*, 22(1), 2013, 172–83.

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

[16] T. Fujimura and T. Toda, "Analysis of noisy-target training for DNN-based speech enhancement", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[17] 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*, 27, 2014.

[18] M. Hashiba, N. Uemi, Y. Yamaguchi, Y. Sugai, and T. Ifukube, "Industrialization of the electrolarynx with a pitch control function and its evaluation", *IEICE Transactions on Information and Systems, D-II*, 94(6), 2001, 1240–7.

[19] 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.

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

[21] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 5901–5.

[22] 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", in *Interspeech*, 2020, 2472–6.

[23] C.-Y. Huang, K.-W. Chang, and H.-Y. Lee, "Toward degradation-robust voice conversion", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6777–81.

[24] C.-Y. Huang, Y. Y. Lin, H.-Y. Lee, and L.-S. Lee, "Defending your voice: Adversarial attack on voice conversion", in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, 552–9.

[25] W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6672–6.

[26] 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*, 29, 2021, 745–55.

[27] 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 *Interspeech*, 2020, 4676–80.

[28] 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 *Interspeech*, 2021, 1329–33.

[29] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations", in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 5944–8.

[30] 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*, 16(6), 2022, 1308–18.

[31] H. Kameoka, K. Tanaka, D. Kwany, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2020, 1849–63.

[32] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 7524–8.

[33] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech", in *International Conference on Machine Learning*, PMLR, 2021, 5530–40.

[34] K. Kobayashi and T. Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on CLDNN", in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, 2018, 2115–9.

[35] K. Kobayashi and T. Toda, "Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN", in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, 396–400.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 626–30.

[37] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", *Advances in Neural Information Processing Systems*, 13, 2000.

[38] J. Li, R. Gadde, B. Ginsburg, and V. Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation", *arXiv preprint arXiv:1811.00707*, 2018.

[39] 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*, 27(8), 2019, 1256–66.

[40] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network", in *Interspeech*, 2018, 342–6.

[41] 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 *46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC)*, IEEE, 2024, 1–4.

[42] 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 *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, 870–7.

[43] 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 *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, 949–54.

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

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

[46] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*, 2018.

[47] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 552–68.

[48] 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*, 10(1), 2019, 151.

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

[50] 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*, 99(7), 2016, 1877–84.

[51] A. Mottini, J. Lorenzo-Trueba, S. V. K. Karlapati, and T. Drugman, "Voicy: Zero-shot non-parallel voice conversion in noisy reverberant environments", in *11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, 113–7.

[52] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech", *Speech Communication*, 54(1), 2012, 134–46.

[53] T.-N. Nguyen, N.-Q. Pham, and A. Waibel, "Accent conversion using pre-trained model and synthesized data from voice conversion", in *Interspeech*, 2022, 2583–7.

[54] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books", in *ICASSP 2015-2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, 5206–10.

[55] 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 *Interspeech*, 2020, 2637–41.

[56] S.-W. Park, D.-Y. Kim, and M.-C. Joe, "Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data", *arXiv preprint arXiv:2005.03295*, 2020.

[57] 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 *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, 546–52.

[58] 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 Radford, "Learning transferable visual models from natural language supervision", in *International Conference on Machine Learning*, PMLR, 2021, 8748–63.

[59] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition", in *Interspeech*, 2019, 3465–9.

[60] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement", in *Interspeech*, 2016, 3743–7.

[61] M. I. Singer and E. D. Blom, "An endoscopic technique for restoration of voice after laryngectomy", *Annals of Otology, Rhinology and Laryngology*, 89(6), 1980, 529–33.

[62] 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.

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

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

[65] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech", in *ICASSP 2010-2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, 4214–7.

[66] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment", in *2012 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2012, 313–7.

[67] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6805–9.

[68] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion", in *Interspeech*, 2013, 3067–71.

[69] C. G. Tang and C. F. Sinclair, "Voice restoration after total laryngectomy", *Otolaryngologic Clinics of North America*, 48(4), 2015, 687–702.

[70] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 15(8), 2007, 2222–35.

[71] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech", in *2016 9th ISCA Speech Synthesis Workshop (SSW)*, 2016, 146–52.

[72] B. Van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge", in *Interspeech*, 2020, 4836–40.

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

[74] L. P. Violeta, W.-C. Huang, D. Ma, R. Yamamoto, K. Kobayashi, and T. Toda, "Electrolaryngeal speech intelligibility enhancement through robust linguistic encoders", in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, 10961–5.

[75] L. P. Violeta, W.-C. Huang, and T. Toda, "Investigating self-supervised pretraining frameworks for pathological speech recognition", in *Interspeech*, 2022, 41–5.

[76] L. P. Violeta, D. Ma, W.-C. Huang, and T. Toda, "Intermediate fine-tuning using imperfect synthetic speech for improving electrolaryngeal speech recognition", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[77] 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*, 32, 2024, 2777–89.

[78] 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 *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 4889–93.

[79] 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*, 9, 2021, 47795–814.

[80] S. E. Williams and J. B. Watson, "Differences in speaking proficiencies in three laryngectomee groups", *Archives of Otolaryngology*, 111(4), 1985, 216–9.

[81] C. Xie and T. Toda, "Noisy-to-Noisy Voice Conversion Under Variations of Noisy Condition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2023, 3871–82.

[82] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6787–91.

[83] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy voice conversion framework with denoising model", in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, 814–20.

[84] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 6199–203.

[85] 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*, 27, 2020, 1730–4.

[86] 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*, 80, 2023, 104279.

[87] 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 *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, 650–7.

[88] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6785–9.

[89] 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 *Interspeech*, 2019, 15–9.

[90] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6790–4.