VITS-Based Data Augmentation for Improved ASR Performance and Domain Adaptation

Anonymous EMNLP submission

Abstract

Although significant advancements have been made in end-to-end speech recognition, it still remains a challenging task when dealing with low-resource scenarios, even with the utilization of traditional data augmentation methods. Recent technological progress, demonstrated by the success of VITS and its variations, has spurred interest in exploring Text-to-Speech (TTS) synthesis for data augmentation to address the aforementioned difficulties. In this study, we investigate the effectiveness of integrating synthetic speech generated by VITS into the train sets of ASR systems. Through comprehensive experiments, we assess the impact of this approach on improving the generalization and performance of ASR models in English, Mandarin, and Japanese. Experimental results indicate that the average character-level accuracy of the VITS-based data augmentation method matches the best performance observed among traditional data augmentation methods before model transfer. After model transfer, the average character-level accuracy of the VITS-based data augmentation method significantly outperforms all traditional methods, surpassing Speed Perturbation, the best-performing traditional method, by 3.5%, as well as Tacotron2 and Fastspeech. Our findings indicate that models trained with the VITS-based data augmentation method exhibit enhanced resilience towards domain shift challenges, demonstrating improved adaptability across varied linguistic contexts, thus highlighting the potential of VITS as a valuable data augmentation technique.

1 Introduction

Automatic Speech Recognition (ASR) tasks play a critical role in enabling human-computer interaction, information retrieval, and the advancement of speech-based applications across diverse domains. The performance of ASR models relies heavily on the quality and quantity of training data. With an ample supply of high-quality training data, both hybrid models (integrating deep neural networks with Hidden Markov models (DNN-HMM)) and end-to-end models (jointly trained neural network systems) demonstrate nearly equivalent performance (Lüscher et al., 2019). However, acquiring large amounts of high-quality labeled speech data tends to be time-consuming and costly. This challenge, particularly pronounced for low-resource languages or specific domains, exacerbates the difficulty of achieving high performance in low-resource tasks, where end-to-end models, compared to hybrid models, notably lag behind (Medennikov et al., 2020). Moreover, trained ASR models often encounter domain shift when transferred to other datasets(Fan et al., 2022; Hidayaturrahman et al., 2023; Chakrabarty et al., 2023). Domain shift occurs when the distribution of data in the target domain, where the model is deployed, differs significantly from that in the source domain, where the model is trained. This difference can lead to a significant decrease in model performance. Addressing domain shift is therefore crucial for ensuring the robustness and generalization ability of machine learning models across diverse and evolving scenarios.

Data augmentation, as an effective method to enhance the diversity of training data through various transformations and expansions, has been widely applied in fields such as computer vision and natural language processing (Pradana et al., 2023; Joshi et al., 2023; Muthumari et al., 2022). In ASR, data augmentation can not only alleviate the problem of insufficient data but also enhance the robustness of the model, especially in coping with different noise environments and speaker variations. Moreover, data augmentation methods can mitigate the performance degradation of models caused by domain shift to some extent.

Common speech data augmentation methods include Noise Augmentation (Ko et al., 2015), Volume Augmentation, Speed Perturbation (Ko et al., 2015), and Specaugment (Park et al., 2019) which involves time and frequency domain masking. Nevertheless, traditional data augmentation methods have limitations, such as being unable to generate entirely new speech patterns and linguistic variations. They rely on manipulating existing audio, which may not fully capture the diversity and complexity of natural speech. In contrast, Text-to-Speech (TTS) methods represented by the VITS model offer a solution by synthesizing diverse and natural-sounding speech from text as a data augmentation technique. This approach expands the range of available speech patterns and linguistic variations beyond what traditional methods can achieve, thus addressing the shortcomings of traditional data augmentation.

In this paper, we compare the character-level accuracy of four traditional data augmentation methods (Noise Augmentation, Volume Augmentation, Speed Perturbation, and SpecAugment) with VITS, Tacotron2, and Fastspeech, three TTS-based data augmentation methods, on ASR tasks at different multiples of train set expansion. To ensure the generality of our experimental results, we evaluate performance across English, Mandarin, and Japanese languages. Specifically, we use the AN4, Ljspeech (Ito and Johnson, 2017) and VCTK datases (Veaux et al., 2016) for English, the THCHS30 (Wang and Zhang, 2015), CSMSC and AISHELL3 (Shi et al., 2020) datasets for Mandarin, the JUST (Kawahara et al., 2000), JVS (Takamichi et al., 2019) and CSJ datasets (Maekawa, 2003) for Japanese. Our findings indicate that the VITS-based data augmentation method achieves comparable performance to traditional methods, Tacotron2, and Fastspeech before migration, and demonstrates superior performance after migration.

The contributions of our paper are as follows: 1) We conduct a comparative analysis between TTS-based data augmentation methods (Tacotron2, Fastspeech, and VITS) and traditional data augmentation methods on multilingual small datasets to simulate tasks with limited resources. This approach aims to enhance the universality of our experimental findings, providing insights into the effectiveness of different augmentation methods. 2) We validate that models trained using VITS-based data augmentation method exhibit superior generalization performance after transfer compared to models trained using other data augmentation methods. This validation underscores the potential of VITS in enhancing model robustness and mitigating domain shift issues, thus contributing to advancements in the field of machine learning and speech processing.

2 Related Work

End-to-End Speech Recognition: End-to-end speech recognition (ASR) has witnessed significant advancements in recent years with models such as Deep Speech (Hannun et al., 2014), Wav2Letter (Collobert et al., 2016), and Listen, Attend and Spell (LAS) (Chan et al., 2015) demonstrating high accuracy in various applications. These models simplify the traditional ASR pipeline by integrating acoustic, language, and pronunciation models into a single neural network.

However, their performance heavily depends on the availability of large amounts of labeled data, which poses a challenge in low-resource scenarios. To address data scarcity, various data augmentation techniques have been employed, such as SpecAugment and Speed Perturbation.

Text-to-Speech (TTS) Synthesis for Data Augmentation: Recent advancements in TTS technology have provided new opportunities for data augmentation in ASR. Models like Tacotron2 (Shen et al., 2017), FastSpeech (Ren et al., 2019), and VITS (Kim et al., 2021) can generate high-quality synthetic speech, which can be used to augment training datasets for ASR systems. Tacotron2 synthesizes natural-sounding speech by converting text into mel-spectrograms and then using a vocoder to generate waveforms. FastSpeech improves on Tacotron2 by using a non-autoregressive approach, making it faster and more stable. VITS combines variational inference with adversarial learning to produce even more natural and diverse speech.

Despite the significant advances in TTS technology over recent years, research on using TTS for data augmentation to address data scarcity in low-resource tasks remains scarce. Within the limited research available, most studies have focused narrowly on the straightforward application of TTS for data augmentation in ASR, without thorough comparisons between TTS as a novel augmentation method and traditional augmentation techniques. For example, Rosenberg et al. evaluated the feasibility of enhancing speech recognition performance by synthesizing speech using two corpora from different domains. Wang et al. explored using text-to-speech data augmentation to enhance children's speech recognition systems. Additionally, there has been insufficient investigation into how the use of these methods affects domain shift issues.

3 Methods

3.1 Traditional Data Augmentation Methods

Speed Perturbation augments the train set by adjusting the playback speed of speech signals. Employing the data augmentation technique, models become more adept at accommodating diverse speaking rates and intonations, consequently bolstering its resilience and generalization capacity. Suppose the original audio signal is $x(t)$, then the signal after Speed Perturbation, $y(t)$, can be expressed as:

$$y(t) = x\left(\frac{t}{\alpha}\right) \tag{1}$$

where α is the Speed Perturbation factor. When $\alpha < 1$, the playback speed of the audio speeds up; when $\alpha > 1$, the playback speed of the audio slows down. In this paper, we set α to 0.9, 1.1, and 1.2, corresponding to accelerating to 111% of the original speed, and slowing down to 91% and 83% of the original speed, respectively.

SpecAugment is a data augmentation technique commonly used in automatic speech recognition (ASR) to enhance model robustness and generalization. It operates by applying three types of transformations to the spectrograms of speech signals: Time Warping (TW), Frequency Masking (FM), and Time Masking (TM). The augmented spectrogram S' is obtained by sequentially applying these transformations to the original spectrogram S:

$$S'(t, f) = \mathrm{TW}(\mathrm{FM}(\mathrm{TM}(S(t, f)))) \tag{2}$$

The formulas for Frequency Masking (FM), Time Masking (TM), and Time Warping (TW) are as follows:

$$\mathrm{TW}(S(t, f)) = S(t + \delta t, f) \tag{3}$$

$$\mathrm{FM}(S(t, f)) = \begin{cases} 0 & \text{if } f_0 \leq f \leq f_0 + \delta f \\ S(t, f) & \text{otherwise} \end{cases} \tag{4}$$

$$\mathrm{TM}(S(t, f)) = \begin{cases} 0 & \text{if } t_0 \leq t \leq t_0 + \delta t \\ S(t, f) & \text{otherwise} \end{cases} \tag{5}$$

In this context, f_0 denotes the random starting frequency for FM, t_0 represents the random starting time for TM, δf stands for the masking width for FM, and δt serves as both the masking width for TM and a random time offset for TW.

Noise Augmentation is also a widely used technique in speech data augmentation. It aims to simulate different noise conditions found in real-world environments, helping to strengthen the model's robustness. In this paper, we implement Noise Augmentation by injecting white noise into the original audio. The amplitude of the added white noise is set to specific proportions of the original audio amplitude. The calculation formula for Noise Augmentation is as follows:

$$y(t) = x(t) + \beta \cdot n(t) \tag{6}$$

where $y(t)$ is the augmented audio signal, $x(t)$ is the original audio signal, $n(t)$ is the white noise signal, β is the scaling factor for the white noise amplitude. In our case, β takes the values 0.01, 0.02, and 0.05.

Volume Augmentation is a data augmentation technique designed to create varied datasets by adjusting the volume of audio signals. By modulating the amplitude of these signals, it simulates recordings with different volume levels, thereby diversifying the training data. The corresponding formulas are given below:

$$y(t) = \gamma \cdot x(t) \tag{7}$$

γ is the scaling factor for the audio signal's amplitude, representing the volume adjustment. In this paper, γ is set to 0.5, 0.9, and 1.1.

Figure 1 illustrates the comparison between the mel spectrogram of the original speech data and the mel spectrograms of the speech data after applying four different traditional data augmentation methods. From left to right in Figure 1 are the original spectrogram, the spectrogram after Speed Perturbation, the spectrogram after SpecAugment, the spectrogram after Noise Augmentation, and the spectrogram after Volume Augmentation. This comparison can provide insights into the effectiveness of these augmentation techniques in enhancing the robustness and generalization capability of the speech recognition model.

3.2 VITS

Our approach leverages the Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS) model. VITS

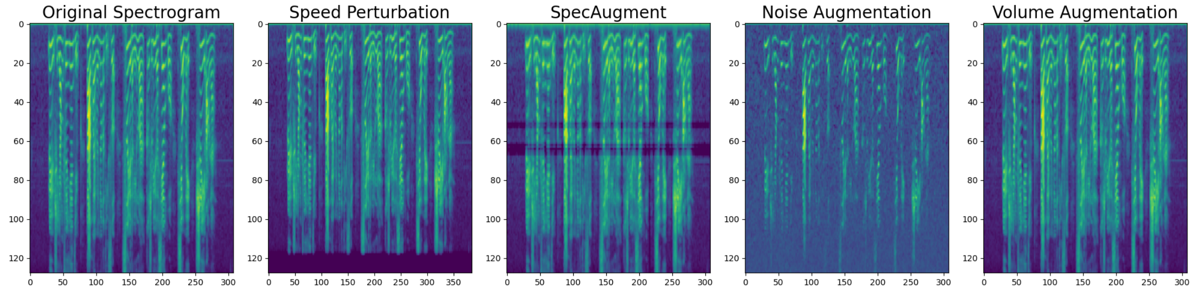

Figure 1: Comparison of Mel Spectrograms: Original Audio and Audio Augmented with Traditional Data Augmentation Methods

is a cutting-edge architecture designed for text-to-speech (TTS) synthesis. It integrates a variational autoencoder (VAE) with adversarial learning techniques to generate high-quality speech waveforms directly from input text. Unlike traditional TTS models, VITS incorporates a conditional mechanism, enabling fine-grained control over the synthesized speech characteristics.

The VITS model consists of an encoder-decoder architecture, where the encoder processes input text into a latent representation that captures the underlying features of the speech. The decoder then generates speech waveforms from this latent representation, producing natural-sounding speech with desired characteristics.

During training, VITS utilizes a combination of loss functions, including spectral loss, duration loss, and adversarial loss, to ensure the synthesized speech's quality and fidelity to the input text. The corresponding equations are shown below.

$$\mathcal{L}_{\text{spec}} = \text{MSE}(S_{\text{synth}}, S_{\text{tar}}) \tag{8}$$

$$\mathcal{L}_{\text{dur}} = \text{MSE}(D_{\text{synth}}, D_{\text{tar}}) \tag{9}$$

$$\mathcal{L}_{\text{adv}} = \log(1 - D(S_{\text{synth}})) - \log(D(S_{\text{tar}})) \tag{10}$$

S_{synth} and S_{tar} represent the synthesized and target speech spectrograms, respectively, in Equation (7). D_{synth} and D_{tar} denote the synthesized and target speech durations, respectively, in Equation (8). $D(S_{\text{synth}})$ and $D(S_{\text{tar}})$ are the discriminator's outputs for the synthesized and target speech spectrograms, respectively, in Equation (9). MSE stands for Mean Squared Error.

VITS offers several advantages, including high-quality speech synthesis, fine-grained control over synthesized speech characteristics, and simplified training and inference procedures. These attributes make it a compelling choice for various TTS applications, ranging from assistive technologies to entertainment and communication systems. The comparison of Mel Spectrograms between the original audio and its corresponding VITS-synthesized audio can be seen in Figure 2.

3.3 ASR

In our approach, we utilized a Transformer-based ASR model architecture, consisting of Conformer encoders and Transformer decoders. The encoder module incorporated Conformer blocks, which combined convolutional and self-attention layers to capture both local and global information from the input speech features. Meanwhile, the decoder module employed Transformer blocks to decode the encoded features and generate the final transcript.

The Conformer block and Transformer decoder can be defined as follows:

$$\text{Conformer} = \text{Conv1D}(\text{SelfAttn}(x)) \tag{11}$$

$$\text{Decoder} = \text{SelfAttn}(h) + \text{FFN}(h) \tag{12}$$

In the provided equations, x denotes the input sequence, which constitutes the raw sequential data processed by the model. h represents the output of the encoder, which serves as the input to the decoder operation. $\text{FFN}()$ refers to the feed-forward neural network. $\text{SelfAttn}()$ represents the operation of applying self-attention mechanism.

During training, the model also optimized a combination of loss functions, comprising spectral loss, duration loss, and adversarial loss, to maintain the accuracy of synthesized speech in relation to the input text.

Optimization was performed using the Adam optimizer with gradient accumulation and gradient clipping to stabilize training. Learning rate scheduling was implemented using the Noam scheduler, gradually increasing the learning rate during the warm-up phase.

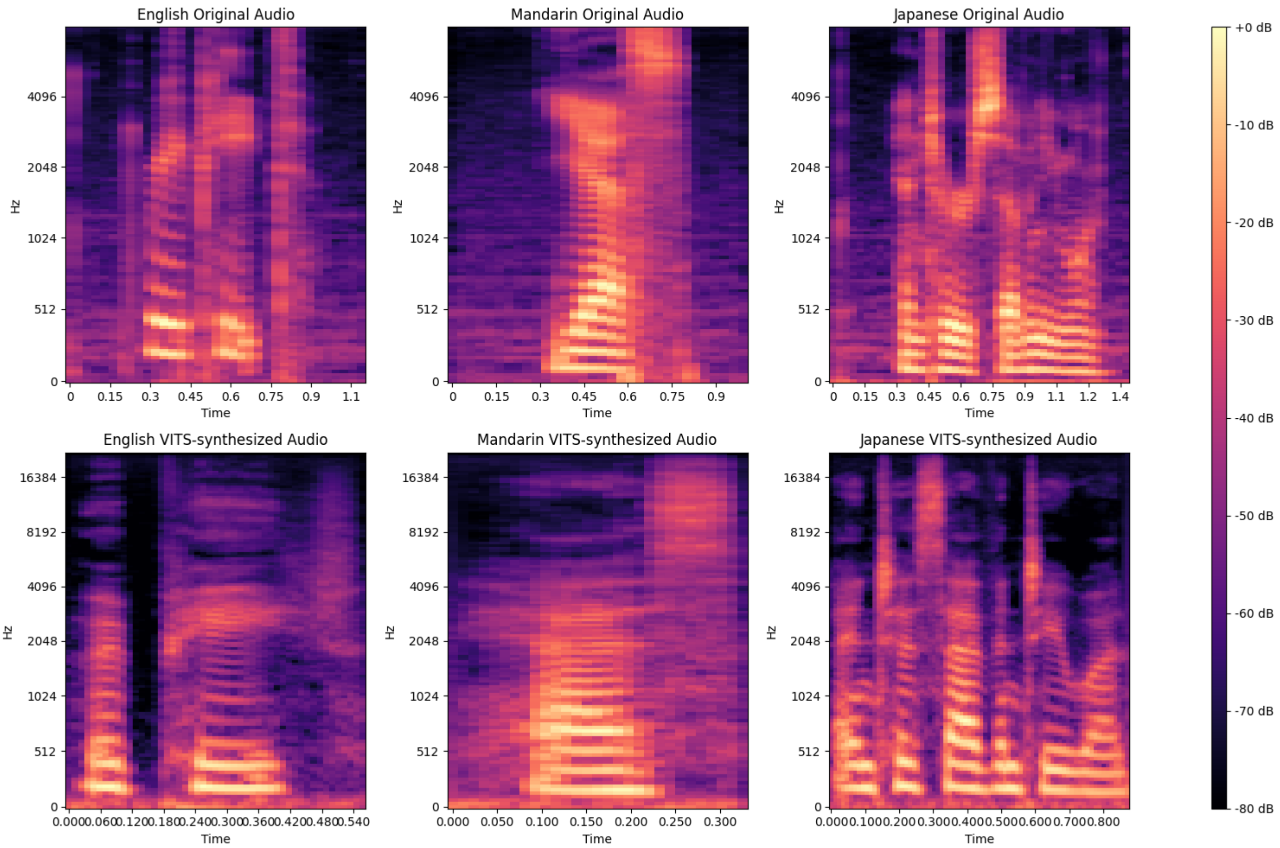

Figure 2: Comparison of Mel Spectrograms: Original Audio and VITS-synthesized Audio

In addition, traditional data augmentation techniques along with VITS-based data augmentation methods were introduced at this stage to augment the training dataset for the ASR task, aiming to enhance the model's performance. These augmentation methods are used to expand the diversity and quantity of training samples, thus providing the model with more varied and representative data to learn from. By exposing the model to a wider range of acoustic variations and perturbations, the augmentation techniques encourage the model to learn more robust and invariant features, thereby improving its ability to generalize to unseen data and challenging acoustic environments.

4 Experiments

In this paper, we train corresponding end-to-end ASR models using the AN4, JSUT, and THCHS30 datasets, and augment the data using synthesized speech obtained from VITS, Tacotron2, and Fastspeech models trained on the VCTK, ASHELL3, and JVS datasets. After migration, the trained ASR models will be tested for their performance on the LJSpeech, CSMSC, and CSJ datasets using newly sampled test sets containing an equivalent number of samples as the original test sets. The details of the train set, dev set, and test set for the aforementioned datasets can be found in Table 1.

During the experiments, we introduce four traditional data augmentation techniques alongside three TTS-based data augmentation methods led by VITS to enhance the train sets. To assess the effectiveness of these augmentation methods, we conduct a comparative analysis of model performance at the character level on the test sets. These evaluations enable us to gauge the models' performance post-migration and investigate the impact of different augmentation techniques on model generalization capabilities. Through meticulous comparisons, we provide evidence supporting the superiority of the VITS-based data augmentation method.

4.1 Datasets

AN4: The AN4 dataset encompasses a diverse collection of speech recordings. With a sample rate of 16000 Hz, it provides a comprehensive resource for developing and evaluating speech recognition algorithms.

LJSpeech: The LJSpeech dataset is a public do-

End-to-end ASR Model Training				End-to-end TTS Model Training				Migration Testing	
Dataset	**train set**	**dev set**	**test set**	**Dataset**	**train set**	**dev set**	**test set**	**Dataset**	**test set**
AN4	878	100	100	VCTK	36000	4000	4000	LJSpeech	100
THCHS30	10708	1340	1340	AISHELL3	68035	10000	10000	CSMSC	1340
JSUT	7196	250	250	JVS	15000	1000	1000	CSJ	250

Table 1: Datasets for ASR Model Training, TTS Model Training, and Migration Testing

main speech dataset consisting of 13,100 short audio clips of a single speaker. Each clip is accompanied by a transcription. The clips vary in length from 1 to 10 seconds, totaling approximately 24 hours of audio. The dataset has a sampling rate of 22.05 kHz.

VCTK: The VCTK Corpus consists of speech data from 110 English speakers with diverse accents. Each speaker reads approximately 400 sentences. All recordings were downsampled to 48 kHz, and manually end-pointed.

THCHS30: The THCHS30 dataset is an open Mandarin speech database. It contains speech data recorded in a quiet office environment, with a total duration exceeding 30 hours. The recordings have a sampling rate of 16 kHz.

CSMSC: CSMSC is a dataset containing approximately 12 hours of effective speech data. The recordings are made by a female speaker aged between 20 to 30 years. The speech data is provided with a sampling rate of 48 kHz.

AISHELL3: The AISHELL-3 dataset is a multi-speaker Mandarin audio corpus. It comprises 88,035 recordings from 218 native speakers reading text from provided scripts with neutral emotions. The recordings were captured at a sampling rate of 44.1 kHz.

JSUT: The JSUT dataset, a novel Japanese speech corpus, comprises approximately 10 hours of speech data recorded by a female native Japanese speaker. It offers speech data in 16-bit/sample RIFF WAV format with a sampling rate of 48 kHz, along with UTF-8 encoded transcriptions of the speech utterances.

CSJ: The Corpus of Spontaneous Japanese (CSJ) is a comprehensive Japanese corpus extensively used for research in phonetics, linguistics, and pragmatics. It comprises approximately 650 hours of spontaneous speech, equivalent to about 7,000k words. The speech materials are recorded using head-worn close-talking microphones and DAT, and downsampled to 16 kHz with 16-bit accuracy. In this experiment, only a portion of the data was utilized for testing purposes.

JVS: The JVS corpus provides high-quality audio recordings from 100 native Japanese speakers, including voice actors and actresses. It features various speech styles such as normal, whispering and falsetto voices, ensuring a diverse dataset. The audio files are sampled at 24 kHz and encoded at 16-bit depth, ensuring excellent fidelity and clarity.

These datasets collectively support the development and evaluation of end-to-end speech recognition models across English, Mandarin, and Japanese languages, enabling a thorough analysis of different data augmentation techniques.

4.2 ASR setup

We utilize the train set, dev set, and test set derived from the AN4, THCHS30, and JSUT datasets for training, model tuning, and evaluation of the ASR models, respectively. The acoustic features, extracted from the raw audio data, consist of 80-dimensional log Mel filterbank coefficients, which have undergone cepstral mean and variance normalization.

Text data is tokenized into subwords using SentencePiece, incorporating special characters "<unknown>", "<blank>", underscore "_", and "<sos/eos>" as sentence boundary markers. For the AN4 English dataset, the vocabulary size is 30. For the THCHS30 Mandarin dataset, it is 2669. For the JSUT Japanese dataset, it is 2742.

The model architecture primarily utilizes the Transformer framework for training the automatic speech recognition model. It employs a folded batch type with a batch size of 64 and sets a maximum of 200 training epochs. Model parameters are initialized using Xavier uniform distribution. The selection criterion for the best model is based on the maximum accuracy achieved on the dev set, retaining the top 10 best models.

Taking into account the linguistic and dataset-specific characteristics, we made slight adjustments

English	1x (%)	2x (%)	3x (%)	4x (%)	4x Post-migration (%)
Raw	86.1	86.4	87.0	86.2	61.9 (-24.3)
Speed Perturbation	86.1	**91.8**	**94.2**	**95.5**	**72.2** (-23.3)
SpecAugment	86.1	**88.2**	**90.5**	92.5	65.9 (-26.6)
Noise Augmentation	86.1	84.8	83.5	83.1	58.2 (-24.9)
Volume Augmentation	86.1	87.5	88.7	88.0	61.6 (-26.4)
Tacotron2	86.1	87.1	89.5	92.4	69.3 (-23.1)
Fastspeech	86.1	86.9	88.1	89.7	65.8 (-23.9)
VITS	86.1	87.8	90.3	**93.4**	**77.1 (-16.3)**
Mandarin	**1x (%)**	**2x (%)**	**3x (%)**	**4x (%)**	**4x Post-migration (%)**
Raw	47.9	48.1	48.2	47.9	43.1 (-4.8)
Speed Perturbation	47.9	**48.6**	**48.8**	49.0	43.7 (-5.3)
SpecAugment	47.9	**48.5**	**48.9**	**49.1**	39.0 (-10.1)
Noise Augmentation	47.9	47.4	47.7	46.6	40.6 (-6)
Volume Augmentation	47.9	46.9	47.5	48.1	42.8 (-5.3)
Tacotron2	47.9	48.3	48.7	47.6	**44.1** (-3.5)
Fastspeech	47.9	48.1	48.5	48.5	43.8 (-4.7)
VITS	47.9	48.0	48.6	**49.3**	**44.8 (-4.5)**
Japanese	**1x (%)**	**2x (%)**	**3x (%)**	**4x (%)**	**4x Post-migration (%)**
Raw	82.4	86.5	86.7	86.5	70.4 (-16.1)
Speed Perturbation	82.4	**86.8**	**87.8**	**88.1**	**72.7** (-15.4)
SpecAugment	82.4	86.5	87.4	87.4	71.6 (-15.8)
Noise Augmentation	82.4	86.6	87.0	87.0	70.2 (-16.8)
Volume Augmentation	82.4	86.2	86.7	87.4	70.7 (-16.7)
Tacotron2	82.4	86.3	87.7	88.0	72.6 (-15.4)
Fastspeech	82.4	85.5	87.1	86.8	71.7 (-15.1)
VITS	82.4	**86.7**	**87.9**	**88.5**	**77.3 (-11.2)**

Table 2: The table illustrates the character-level accuracy of ASR models on the test sets, achieved by expanding train sets using different data augmentation methods to various multiples. The labels 1x, 2x, 3x, and 4x denote the expansion multiples of the train sets. 4x Post-migration represents the accuracy of the ASR models after transfer, when the train sets have been expanded to four times its original size.

to the structures of the ASR models for the three languages, aiming to achieve optimal performance. For the end-to-end ASR model trained on the AN4 dataset and THCHS30 dataset, the encoder produces a 256-dimensional output, with 4 attention heads per mechanism and 12 Transformer blocks. Similarly, the decoder includes 4 attention heads per mechanism and 6 Transformer blocks. The model uses a hybrid loss function that combines spectral loss, duration loss, and adversarial loss, each with specific weights. It adopts the Adam optimizer with a learning rate of 0.001 and employs a warm-up learning rate scheduler with 2500 warm-up steps. During the inference phase, the beam search algorithm is used, with a beam size set to 10.

For the end-to-end ASR model trained on the JSUT dataset, the encoder consists of 12 blocks with 2048 linear units, 4 attention heads per mechanism, and a 256-dimensional output. It incorporates dropout with a rate of 0.1 and utilizes a swish activation function. Additionally, it includes relative position encoding and self-attention mechanisms, as well as a convolutional neural network (CNN) module with a kernel size of 15. The decoder comprises 6 transformer blocks with 2048 linear units and a dropout rate of 0.1. The beam size for the beam search algorithm during the inference phase is set to 20.

4.3 TTS setup

In our experimental setup for Text-to-Speech (TTS), we leveraged the advanced capabilities of the VITS (Variational Inference Transformer for Speech Synthesis) model. This state-of-the-art model, renowned for its remarkable performance in generating natural-sounding speech, formed the cornerstone of our TTS experiments.

For the end-to-end TTS models in three languages, we loaded pre-trained multi-speaker TTS model parameters trained on Aishell3, VCTK, and JVS datasets. By leveraging these pre-trained models, we synthesized speech based on the text from the train sets of the ASR task, thereby augmenting the scale of the training data and enhancing diversity.

4.4 Comparison

In our experiments, we expand the train set to double, triple, and quadruple its original size using different data augmentation techniques. Subsequently, we evaluate the performance of the trained ASR models on the corresponding test sets in terms of character-level accuracy.

As shown in Table 2, introducing either the TTS-based data augmentation method or traditional data augmentation techniques can enhance the performance of the trained models in English, Mandarin, and Japanese. In the scenarios of Mandarin and Japanese, the ASR models trained using the VITS-based data augmentation method achieved the highest accuracy. In the scenario where the training datasets for all three languages are quadrupled, the accuracy of ASR models trained using the VITS-based data augmentation method consistently ranks within the top two. The average accuracy across the three scenarios is 77%. This performance is comparable to that of the Speed Perturbation and SpecAugment methods and significantly surpasses the Noise Augmentation and Volume Augmentation methods, with an average accuracy improvement of 4.8% and 2.5%, respectively, across all three scenarios. Besides, the performance of the VITS-based data augmentation method outperforms Tacotron2 and Fastspeech, both of which are also TTS-based data augmentation methods.

Due to differences in data distribution between datasets, often referred to as the domain shift problem, trained models tend to suffer performance degradation when transferred to new datasets. In our experiments, we transferred the ASR models trained with various data augmentation methods to new datasets—LJSpeech, CSMSC, and CSJ—and tested their character-level accuracy on these new datasets. From Table 2, it can be observed that under the condition of a fourfold expansion of the train set in three languages, the performance degradation of the ASR models trained using the VITS-based data augmentation method is the least when transferred to new datasets, averaging a decrease of 10.7%. Conversely, among other data augmentation methods, Speed Perturbation exhibits the best performance post-migration, with an average decrease of 14.7% across the three scenarios. Following model transfer, the ASR models trained using the VITS-based data augmentation method consistently achieve the highest accuracy in all three scenarios, with an average character-level accuracy of 66.4%, surpassing Speed Perturbation, the second-best performer, by approximately 3.5%.

Overall, ASR models trained using the VITS-based data augmentation method outperform most data augmentation methods in terms of performance before model transfer, slightly outpacing SpecAugment and Speed Perturbation. After model transfer, the character-level accuracy of ASR models trained using the VITS-based data augmentation method significantly surpasses that of other data augmentation methods in all scenarios, with the smallest decrease in accuracy compared to other methods.

These experimental results clearly demonstrate the advantages of the VITS-based data augmentation method over traditional data augmentation methods. Additionally, they also confirm the superiority of the VITS-based data augmentation method among TTS-based data augmentation methods.

5 Conclusion and Future Work

Through comprehensive experiments, we demonstrate that integrating synthetic speech generated by VITS-based data augmentation into ASR train sets significantly improves the performance and generalization of ASR systems in low-resource scenarios. Moreover, ASR models trained with VITS-based data augmentation exhibit enhanced resilience to domain shifts and better adaptability across various linguistic contexts compared to the remaining data augmentation methods. Our study underscores the potential of TTS represented by VITS as a valuable data augmentation method, offering a practical solution to the challenges faced by ASR systems in low-resource scenarios. Future research could concentrate on advancing VITS-based TTS models to maximize augmentation benefits by enhancing fidelity, adaptability across languages, and scalability to broader ASR applications.

Limitations

Despite the promising results obtained, several limitations should be acknowledged. First, the evaluation was conducted on a limited set of languages (English, Mandarin, and Japanese), potentially limiting the generalizability of findings to other low-resource languages with different phonetic structures or linguistic characteristics, for example, Tibetan. Second, the impact of hyperparameter settings and specific implementation details of the VITS-based augmentation method were not extensively explored, which could influence its effectiveness across different datasets and tasks. Finally, while improvements in model robustness were observed during testing, longitudinal studies to assess the long-term stability of models under varying conditions are warranted. Further in-depth research is needed to fully understand and address these limitations, and the path ahead requires continuous effort and dedication.

Ethical Statement

This study adheres strictly to ethical principles, ensuring the confidentiality of data and ethical conduct in all research practices. The research rigorously upholds the principles of beneficence, non-maleficence, and justice. All data utilized in this study are sourced exclusively from publicly available datasets.

References

Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. 2023. A simple signal for domain shift. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3569–3576.

William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. Listen, attend and spell. *CoRR*, abs/1508.01211.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. 2016. Wav2letter: an end-to-end convnet-based speech recognition system. *CoRR*, abs/1609.03193.

Jiahao Fan, Hangyu Zhu, Xinyu Jiang, Long Meng, Chen Chen, Cong Fu, Huan Yu, Chenyun Dai, and Wei Chen. 2022. Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:205–216.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

Hidayaturrahman, Yaya Heryadi, Lukas, Wayan Suparta, and Yulyani Arifin. 2023. Exploring shifted domain problem within mnist dataset. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pages 750–754.

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

Deepali Joshi, Aryan Shinde, Shreya Das, Om Deokar, Dipasha Shetiya, and Simran Jagtap. 2023. Text data augmentation. In *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, pages 392–396.

Tatsuya Kawahara, Akinobu Lee, Tetsunori Kobayashi, Kazuya Takeda, Nobuaki Minematsu, Shigeki Sagayama, Katunobu Itou, Akinori Ito, Mikio Yamamoto, Atsushi Yamada, Takehito Utsuro, and Kiyohiro Shikano. 2000. Free software toolkit for japanese large vocabulary continuous speech recognition. volume 4, pages 476–479.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *CoRR*, abs/2106.06103.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech*.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. Rwth asr systems for librispeech: Hybrid vs attention. pages 231–235.

Kikuo Maekawa. 2003. Corpus of spontaneous japanese: Its design and evaluation. *Proceedings of SSPR*.

Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko. 2020. The STC System for the CHiME-6 Challenge. In *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pages 36–41.

M. Muthumari, C Ambika Bhuvaneswari, J Eswar Naga Sai Kumar Babu, and S Prudhvi Raju. 2022. Data augmentation model for audio signal extraction. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 334–340.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*.

Rilo Chandra Pradana, Setiawan Joddy, and Abba Suganda Girsang. 2023. Easy data augmentation for handling imbalanced data in fake news detection. In *2023 International Conference on Technology, Engineering, and Computing Applications (ICTECA)*, pages 1–5.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *CoRR*, abs/1905.09263.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro J. Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech. *CoRR*, abs/1909.11699.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2017. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines. *CoRR*, abs/2010.11567.

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. 2019. JVS corpus: free japanese multi-speaker voice corpus. *CoRR*, abs/1908.06248.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2016. Superseded - cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Dong Wang and Xuewei Zhang. 2015. THCHS-30 : A free chinese speech corpus. *CoRR*, abs/1512.01882.

Wei Wang, Zhikai Zhou, Yizhou Lu, Hongji Wang, Chenpeng Du, and Yanmin Qian. 2021. Towards data selection on tts data for children's speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6888–6892.