The NPU-MSXF Speech-to-Speech Translation System for IWSLT 2023 Speech-to-Speech Translation Task

Kun Song[1], **Yi Lei**[1], **Peikun Chen**[1], **Yiqing Cao**[2], **Kun Wei**[1], **Yongmao Zhang**[1], **Lei Xie**[1*], **Ning Jiang**[3], **Guoqing Zhao**[3]

[1]Audio, Speech and Language Processing Group (ASLP@NPU),
School of Computer Science, Northwestern Polytechnical University, China
[2]Department of Computer Science and Technology, Nanjing University, China
[3]MaShang Consumer Finance Co., Ltd, China

Abstract

This paper describes the NPU-MSXF system for the IWSLT 2023 speech-to-speech translation (S2ST) task which aims to translate from English speech of multi-source to Chinese speech. The system is built in a cascaded manner consisting of automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). We make tremendous efforts to handle the challenging multi-source input. Specifically, to improve the robustness to multi-source speech input, we adopt various data augmentation strategies and a ROVER-based score fusion on multiple ASR model outputs. To better handle the noisy ASR transcripts, we introduce a three-stage fine-tuning strategy to improve translation accuracy. Finally, we build a TTS model with high naturalness and sound quality, which leverages a two-stage framework, using network bottleneck features as a robust intermediate representation for speaker timbre and linguistic content disentanglement. Based on the two-stage framework, pre-trained speaker embedding is leveraged as a condition to transfer the speaker timbre in the source English speech to the translated Chinese speech. Experimental results show that our system has high translation accuracy, speech naturalness, sound quality, and speaker similarity. Moreover, it shows good robustness to multi-source data.

1 Introduction

In this paper, we describe NPU-MSXF team's cascaded speech-to-speech translation (S2ST) system submitted to the speech-to-speech (S2S) track[1] of the IWSLT 2023 evaluation campaign. The S2S track aims to build an offline system that realizes speech-to-speech translation from English to Chinese. Particularly, the track allows the use of large-scale data, including the data provided in this track as well as all training data from the offline track[2] on speech-to-text translation task. Challengingly, the test set contains multi-source speech data, covering a variety of acoustic conditions and speaking styles, designed to examine the robustness of the S2ST system. Moreover, speaker identities conveyed in the diverse multi-source speech test data are unseen during training, which is called *zero-shot S2ST* and better meets the demands of real-world applications.

Current mainstream S2ST models usually include *cascaded* and *end-to-end* systems. Cascaded S2ST systems, widely used in the speech-to-speech translation task (Nakamura et al., 2006), usually contain three modules, i.e. automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). Meanwhile, end-to-end (E2E) S2ST systems (Jia et al., 2019; Lee et al., 2022) have recently come to the stage by integrating the above modules into a unified model for directly synthesizing target language speech translated from the source language. E2E S2ST systems can effectively simplify the overall pipeline and alleviate possible error propagation. Cascaded S2ST systems may also alleviate the error propagation problem by leveraging the ASR outputs for MT model fine-tuning. Meanwhile, thanks to the individual training process of sub-modules, cascaded systems can make better use of large-scale text and speech data, which can significantly promote the performance of each module.

In this paper, we build a cascaded S2ST system aiming at English-to-Chinese speech translation with preserving the speaker timbre of the source English speech. The proposed system consists of Conformer-based (Gulati et al., 2020) ASR models, a pretrain-finetune schema-based MT model (Radford et al., 2018), and a VITS-based TTS model (Kim et al., 2021). For ASR, model fusion and data augmentation strategies are adopted to improve the recognition accuracy and generalization ability of ASR with multi-source input.

[*] Lei Xie is the corresponding author.

[1] https://iwslt.org/2023/s2s

[2] https://iwslt.org/2023/offline

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 311–320
July 13-14, 2023 ©2023 Association for Computational Linguistics

For MT, we use a three-stage fine-tuning process to adapt the translation model to better facilitate the output of ASR. Meanwhile, back translation and multi-fold verification strategies are adopted. Our TTS module is composed of a text-to-BN stage and a BN-to-speech stage, where speaker-independent neural bottleneck (BN) features are utilized as an intermediate representation bridging the two stages. Specifically, the BN-to-speech module, conditioned on speaker embedding extracted from the source speech, is to synthesize target language speech with preserving the speaker timbre. Combined with a pre-trained speaker encoder to provide speaker embeddings, the TTS model can be generalized to unseen speakers, who are not involved in the training process. Experimental results demonstrate the proposed S2ST system achieves good speech intelligibility, naturalness, sound quality, and speaker similarity.

2 Automatic Speech Recognition

Our ASR module employs multiple models for score fusion in the inference. Moreover, data augmentation is adopted during training to handle noisy multi-source speech.

2.1 Model Structure

Our system employs both Conformer (Gulati et al., 2020) and E-Branchformer models (Kim et al., 2023) in our ASR module to address the diversity of the test set. Conformer sequentially combines convolution, self-attention, and feed-forward layers. The self-attention module serves to capture global contextual information from the input speech, while the convolution layer focuses on extracting local correlations. This model has demonstrated remarkable performance in ASR tasks with the ability to capture local and global information from input speech signals. E-Branchformer uses dedicated branches of convolution and self-attention based on the Conformer and applies efficient merging methods, in addition to stacking point-wise modules. E-Branchformer achieves state-of-the-art results in ASR.

2.2 Data Augmentation

Considering the diversity of the testing data, we leverage a variety of data augmentation strategies to expand the training data of our ASR system, including the following aspects.

- **Speed Perturbation**: We notice that the testing set contains spontaneous speech such as conversations with various speaking speeds. So speed perturbation is adopted to improve the generalization ability of the proposed model. Speed perturbation is the process of changing the speed of an audio signal while preserving other information (including pitch) in the audio. We perturb the audio speech with a speed factor of 0.9, 1.0, and 1.1 to all the training data. Here speed factor refers to the ratio compared to the original speed of speech.
- **Pitch Shifting**: Pitch shifting can effectively vary the speaker identities to increase data diversity. Specifically, we use SoX[3] audio manipulation tool to perturb the pitch in the range [-40, 40].
- **Noise Augmentation**: There are many cases with heavy background noise in the test set, including interfering speakers and music. However, the data set provided by the organizer is much cleaner than the test set, which makes it necessary to augment the training data by adding noises to improve the recognition performance. Since there is no noise set available, we create a noise set from the data provided. A statistical VAD (Sohn et al., 1999) is used to cut the non-vocal and vocal segments from the data and the non-vocal segments with energy beyond a threshold comprise our noise set. We add the noise segments to the speech utterances with a signal-to-noise ratio ranging from 0 to 15 dB.
- **Audio Codec**: Considering the test data come from multiple sources, we further adopt audio codec augmentation to the training data. Specifically, we use the FFmpeg[4] tool to convert the original audio to Opus format at [48, 96, 256] Kbps.
- **Spectrum Augmentation**: To prevent the ASR model from over-fitting, we apply the SpecAugment method (Park et al., 2019) to the input features during every mini-batch training. SpecAugment includes time warping, frequency channel masking, and time step masking, and we utilize all of these techniques during training.

2.3 Model Fusion

Since a single ASR model may overfit to a specific optimization direction during training, it cannot guarantee good recognition accuracy for the

[3] https://sox.sourceforge.net/

[4] https://ffmpeg.org/

speech of various data distributions. To let the ASR model generalize better to the multi-source input, we adopt a model fusion strategy. Specifically, we train the Conformer and E-branchformer models introduced in Section 2.1 using the combination of the original and the augmented data. Each testing utterance is then transcribed by these different models, resulting in multiple outputs. Finally, ROVER (Fiscus, 1997) is adopted to align and vote with equal weights on the multiple outputs, resulting in the final ASR output.

2.4 ASR Output Post-processing

Given that the spontaneous speech in the test set contains frequent filler words such as "Uh" and "you know", it is necessary to address their impact on subsequent MT accuracy and TTS systems that rely on the ASR output. To mitigate this issue, we use a simple rule-based post-processing step to detect and eliminate these expressions from the ASR output. By doing so, we improve the accuracy of the downstream modules.

3 Machine Translation

For the MT module, we first use a pre-trained language model as a basis for initialization and then employ various methods to further enhance translation accuracy.

3.1 Pre-trained Language Model

As pre-trained language models are considered part of the training data in the offline track and can be used in the S2ST track, we use the pre-trained mBART50 model for initializing our MT module. mBART50 (Liu et al., 2020) is a multilingual BART (Lewis et al., 2020) model with 12 layers of encoder and decoder, which we believe will provide a solid basis for improving translation accuracy.

3.2 Three-stage Fine-tuning based on Curriculum Learning

We perform fine-tuning on the pre-trained model to match the English-to-Chinese (En2Zh) translation task. There are substantial differences between the ASR outputs and the texts of MT data. First, ASR prediction results inevitably contain errors. Second, ASR outputs are normalized text without punctuation. Therefore, directly fine-tuning the pre-trained model with the MT data will cause a mismatch problem with the ASR output during inference. On the other hand, fine-tuning the model with the ASR outputs will cause difficulty in model coverage because of the difference between the ASR outputs and the MT data. Therefore, based on Curriculum Learning (Bengio et al., 2009), we adopt a three-stage fine-tuning strategy to mitigate such a mismatch.

- **Fine-tuning using the MT data**: First, we use all the MT data to fine-tune the pre-trained model to improve the accuracy of the model in the En2Zh translation task.

- **Fine-tuning using the MT data in ASR transcription format**: Second, we convert the English text in the MT data into the ASR transcription format. Then, we fine-tune the MT model using the converted data, which is closer to the actual text than the ASR recognition output. This approach can enhance the stability of the fine-tuning process, minimize the impact of ASR recognition issues on the translation model, and improve the model's ability to learn punctuation, thereby enhancing its robustness.

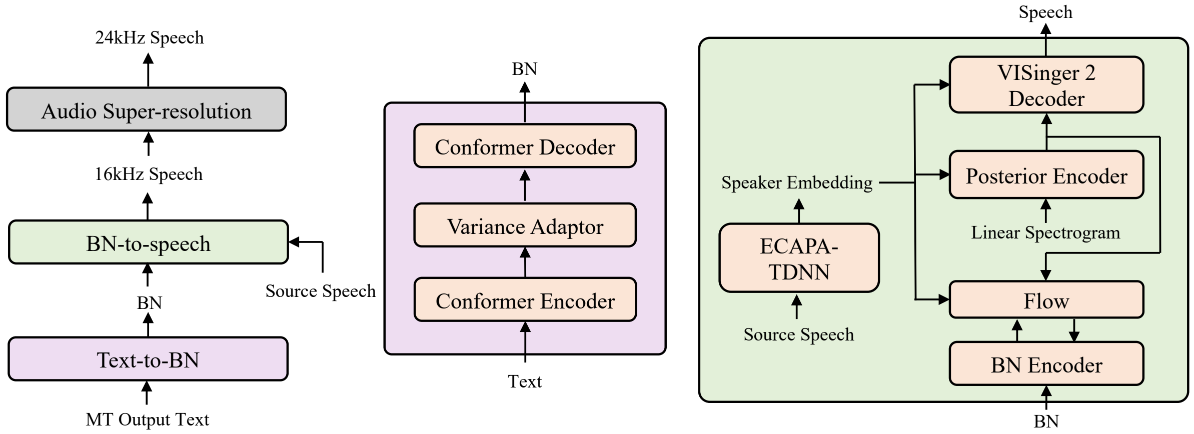
- **Fine-tuning using the ASR outputs**: Third, we leverage *GigaSpeech* (Chen et al., 2021) to address the mismatch problem between the ASR outputs and the MT data. Specifically, we use the ASR module to transcribe the *GigaSpeech* training set and replace the corresponding transcriptions in *GigaST* (Ye et al., 2022) with the ASR transcriptions for translation model fine-tuning. This enables the MT model to adapt to ASR errors.

3.3 Back Translation

Following (Akhbardeh et al., 2021), we adopt the back translation method to enhance the data and improve the robustness and generalization of the model. First, we train a Zh2En MT model to translate Chinese to English, using the same method employed for the En2Zh MT module. Next, we generate the corresponding English translations for the Chinese text of the translation data. Finally, we combine the back translation parallel corpus pairs with the real parallel pairs and train the MT model.

3.4 Cross-validation

We use 5-fold cross-validation (Ojala and Garriga, 2010) to improve the robustness of translation and reduce over-fitting. Firstly, we randomly divide the data into five equal parts and train five models on different datasets by using one of them as the validation set each time and combining the remaining four as the training set. After that, we integrate the predicted probability distributions from these five



(a) pipeline (b) Text-to-BN (c) BN-to-speech

Figure 1: Architecture of our text-to-speech module.

models to obtain the final predicted probability distribution for the next word during token generation for predicting the translation results.

4 Text-to-speech

4.1 Overview

Figure 1 (a) shows the pipeline of the text-to-speech module in the proposed S2ST system. The TTS module is built on a BN-based two-stage architecture, which consists of a text-to-BN and a BN-to-speech procedure. The text-to-BN stage tends to generate BN features from the Chinese text translated by the MT module. The BN-to-speech stage produces 16KHz Chinese speech from the BN feature, conditioning on the speaker embedding of source speech. Given the translated Chinese speech which preserves the speaker timbre in the source English speech, an audio super-resolution model is further leveraged to convert the synthesized speech from 16KHz to 24KHz for higher speech fidelity.

Building on the two-stage framework AdaVITS (Song et al., 2022a), we employ bottleneck (BN) features as the intermediate representations in the two-stage TTS module. BN features, extracted from a multi-condition trained noise-robust ASR system, mainly represent the speaker-independent linguistic content. So BN can effectively *disentangle* the speaker timbre and the linguistic content information. In the text-to-BN stage, high-quality TTS data is adopted in the training phase to model the speaker-independent BN features with prosody information. In the BN-to-speech stage, both high-quality TTS data and low-quality ASR data should be involved during training to sufficiently model the speech of various speaker identities. Extracted from speech, BN features contain the duration and prosody information, which eliminates the need for text transcripts and prosody modeling. Instead, the BN-to-speech stage focuses on time-invariant information modeling, such as speaker timbre.

As the goal of this work is to conduct zero-shot English-to-Chinese speech translation, we concentrate on the method to transfer the unseen speaker timbre of the source English speech to the synthesized Chinese speech through voice cloning (Chen et al., 2019). To capture new speaker timbre during inference, the TTS module requires to model abundant various speakers during training, which relies on large-scale high-quality TTS data. Unfortunately, we are limited in the high-quality TTS data we can use in this task and must rely on additional data such as ASR to model the speaker timbre. However, this data is not suitable for TTS model training because the labels are inconsistent with TTS, and the prosody of the speakers is not as good as high-quality TTS data.

Furthermore, we incorporate ASR data into the BN-to-speech training procedure by re-sampling all the training speech to 16kHz, which can not reach high-quality audio. Therefore, we utilize audio super-resolution techniques to upsample the synthesized 16KHz audio and convert it into higher sampling rate audio.

4.2 Text-to-BN

Our text-to-BN stage network in TTS is based on DelightfulTTS (Liu et al., 2021), which employs a Conformer-based encoder, decoder, and a variance adapter for modeling duration and prosody. The model extends phoneme-level linguistic features to frame-level to guarantee the clarity and naturalness of speech in our system.

4.3 BN-to-speech

We build the BN-to-speech model based on VITS (Kim et al., 2021), which is a mainstream end-to-end TTS model. VITS generates speech waveforms directly from the input textual information, rather than a conventional pipeline of using the combination of an acoustic model and a neural vocoder.

The network of the BN-to-speech stage consists of a BN encoder, posterior encoder, decoder, flow, and speaker encoder. The monotonic alignment search (MAS) from the original VITS is removed since BN features contain the duration information. For achieving zero-shot voice cloning, an ECAPA-TDNN (Desplanques et al., 2020) speaker encoder is pre-trained to provide the speaker embedding as the condition of the synthesized speech. To avoid periodic signal prediction errors in the original HiFiGAN-based (Kong et al., 2020) decoder in VITS, which induces sound quality degradation, we follow VISinger2 (Zhang et al., 2022) to adopt a decoder with the sine excitation signals. Since The VISinger2 decoder requires pitch information as input, we utilize a pitch predictor with a multi-layer Conv1D that predicts the speaker-dependent pitch from BN and speaker embedding. With the desired speaker embedding and corresponding BN features, the BN-to-speech module produces Chinese speech in the target timbre.

4.4 Audio Super-resolution

Following (Liu et al., 2021), we use an upsampling network based vocoder to achieve audio super-resolution (16kHz→24kHz). During training, the 16KHz mel-spectrogram is used as the condition to predict the 24KHz audio in the audio super-resolution model. Specifically, we adopt the *AISHELL-3* (Shi et al., 2021) dataset, composing the paired 16KHz and 24KHz speech data for model training. During inference, the high-quality 24kHz speech is produced for the mel-spectrogram of the 16KHz speech generated by the BN-to-speech model. Here DSPGAN (Song et al., 2022b) is adopted as our audio super-resolution model, which is a universal vocoder that ensures robustness and good sound quality without periodic signal errors.

5 Data Preparation

5.1 Datasets

Following the constraint of data usage, the training dataset for the S2ST system is illustrated in Table 1.

5.1.1 ASR Data

For the English ASR module in our proposed system, we use *GigaSpeech*, *LibriSpeech*, *TED-LIUM v2&v3* as training data. For the ASR system used to extract BN features in TTS, we use text-to-speech data in *AISHELL-3* and Chinese speech in *GigaS2S*, along with the corresponding Chinese text in *GigaST*, as the training set. Since the test set's MT output text is a mix of Chinese and English, including names of people and places, the TTS module needs to support both languages. Therefore, we also add the aforementioned English data to the training set.

5.1.2 MT Data

We use the text-parallel data including *News Commentary* and *OpenSubtitles2018* as MT training set. Moreover, we also add the Chinese texts in *GigaST* and the English texts in *GigaSpeech* corresponding to the Chinese texts in *GigaST* to the training set.

5.1.3 TTS Data

We use *AISHELL-3* as training data in Text-to-BN and audio super-resolution. For the pre-trained speaker encoder, we adopt *LibriSpeech*, which contains 1166 speakers, as the training data.For the BN-to-speech model, in addition to using *AISHELL-3* which has 218 speakers, we also use *LibriSpeech* to meet the data amount and speaker number requirements of zero-shot TTS.

5.2 Data Pre-processing

5.2.1 ASR Data

To prepare the ASR data, we pre-process all transcripts to remove audio-related tags. Next, we map the text to the corresponding byte-pair encoding (BPE) unit and count the number of BPE units in the ASR dictionary, which totals 5,000 units. For audio processing, we use a frame shift of 10ms and a frame length of 25ms and normalize all audio to 16KHz.

5.2.2 MT Data

For the MT data, we use the same tokenizer as mBART50 to perform sub-word segmentation for English and Chinese texts and to organize them into a format for neural network training. By doing so, we can maximize the benefits of initializing our translation model with mBART50 pre-trained model parameters. The mBART tokenizer mentioned above is a Unigram tokenizer. A Unigram model is a type of language model that considers each token to be independent of the tokens before it. What's more, the tokenizer has a total of 250,054 word segmentations, supports word segmentation processing for English, Chinese, and

[5] https://github.com/SpeechTranslation/GigaS2S

Table 1: Datasets used in our proposed system.

Datasets	Utterances	Hours
English Labeled Speech Data		
GigaSpeech (Chen et al., 2021)	8,315K	10,000
LibriSpeech (Panayotov et al., 2015)	281K	961
TED-LIUM v2 (Rousseau et al., 2012)&v3 (Hernandez et al., 2018)	361K	661
CommonVoice (Ardila et al., 2020)	1,225K	1,668
Text-parallel Data		
News Commentary (Chen et al., 2021)	322K	-
OpenSubtitles2018 (Lison et al., 2018)	10M	-
ST Data		
GigaST (Ye et al., 2022)	7,651K	9,781
S2S Data		
GigaS2S[5]	7,626K	-
Chinese TTS Data		
AISHELL-3 (Shi et al., 2021)	88K	85

other languages, and uses special tokens like <s>, </s>, and <unk>.

5.2.3 TTS Data

For *AISHELL-3*, we downsample it to 16KHz and 24KHz respectively as the TTS modeling target and the audio super-resolution modeling target. All other data is down-sampled to 16KHz. All data in TTS adopts 12.5ms frame shift and 50ms frame length.

Speech Enhancement. Given the presence of substantial background noise in the test set, the discriminative power of speaker embeddings is significantly reduced, thereby impeding the performance of the TTS module. Furthermore, the ASR data incorporated during the training of the BN-to-speech model is also subject to background noise. Therefore, we employ a single-channel wiener filtering method (Lim and Oppenheim, 1979) to remove such noise from these data. Please note that we do not perform speech enhancement on the test set in the ASR module, because there is a mismatch between the denoised audio and which is used in ASR training, and denoising will reduce the speech recognition accuracy.

5.2.4 Evaluation Data

For all evaluations, we use the English-Chinese (En-Zh) development data divided by the organizer from *GigaSpeech*, *GigaST* and *GigaS2S*, including 5,715 parallel En-Zh audio segments, and their corresponding En-Zh texts. It is worth noting that the development data for evaluations has been removed from the training dataset.

6 Experiments

6.1 Experimental Setup

All the models in our system are trained on 8 A100 GPUs and optimized with Adam (Kingma and Ba, 2015).

ASR Module. All ASR models are implemented in ESPnet[6]. Both Conformer and E-Branchformer models employ an encoder with 17 layers and a feature dimension of 512, with 8 heads in the self-attention mechanism and an intermediate hidden dimension of 2048 for the FFN. In addition, we employ a 6-layer Transformer decoder with the same feature hidden dimension as the encoder. The E-Branchformer model uses a cgMLP with an intermediate hidden dimension of 3072. The total number of parameters for the Conformer and E-Branchformer model in Section 2.1 is 147.8M and 148.9M respectively. We train the models with batch size 32 sentences per GPU for 40 epochs, and set the learning rate to 0.0015, the warm-up step to 25K.

For data augmentation, we conduct speed perturbation, pitch shifting, and audio codec on the original recordings. Spectrum augmentation and

[6] https://github.com/espnet/espnet

noise augmentation are used for on-the-fly model training.

MT Module. All MT models are implemented in HuggingFace[7]. Using MT data, we fine-tune the mBART-50 large model, which has 611M parameters, with a batch size of 32 sentences per GPU for 20 epochs. The learning rate is set to 3e-5 and warmed up for the first 10% of updates and linearly decayed for the following updates. For fine-tuning using the MT data in ASR transcription format and the ASR outputs, we also fine-tune the model with batch size 32 sentences per GPU for 5 epochs and set the learning rate to 3e-5, which is warmed up for the first 5% of updates and linearly decayed for the following updates.

TTS Module. We complete our system based on VITS official code[8]. The text-to-BN follows the configuration of DelightfulTTS and has about 64M parameters. To extract the duration required for text-to-BN, we train a Kaldi[9] model using *AISHELL-3*. The ASR system used for extracting BN is the Chinese-English ASR model mentioned in Section 5.1.1. For BN-to-speech, we use a 6-layer FFT as the BN encoder and follow the other configuration in VIsinger2 with about 45M parameters in total. The pitch predictor has 4 layers of Conv1D with 256 channels. Pitch is extracted by Visinger2 decoder and DSPGAN from Harvest (Morise, 2017) with Stonemask. To predict pitch in DSPGAN, we use the method described in Section 4.3. Up-sampling factors in DSPGAN is set as [5, 5, 4, 3] and other configuration of DSPGAN-mm is preserved for audio super-resolution. The DSPGAN model has about 9M parameters in total. We train all the above models with a batch size of 64 sentences per GPU for 1M steps and set the learning rate to 2e-4. For the pre-trained speaker encoder, we follow the model configuration and training setup of ECAPA-TDNN (C=1024) with 14.7M parameters.

6.2 Evaluation Models

Baseline. To evaluate the effectiveness of the proposed cascaded S2ST system, we adopt the original cascaded S2ST system as a baseline, including an E-Branchformer ASR model, a mBART50 MT model fine-tuned using the MT data, and an end-to-end TTS model based on VITS trained with *AISHELL-3*.

Proposed system & Ablation Study. We further conduct ablation studies to evaluate each component in the proposed system. Specifically, the ablation studies are designed to verify the effectiveness of model fusion and data augmentation in ASR, three-stage fine-tuning, back translation, cross-verification in MT, two-stage training with BN, pre-trained speaker embedding, and audio super-resolution in TTS.

6.3 Results & Analysis

We conduct experiments on the effectiveness of each sub-module and the performance of our proposed cascaded S2ST system.

6.3.1 ASR Module

We calculate the word error rate (WER) of each ASR module to evaluate the English speech recognition accuracy. As shown in Table 2, the WER of the proposed system has a significant drop compared with the baseline, which indicates that the proposed system greatly improves the recognition accuracy. Moreover, the results of the ablation study demonstrate the effectiveness of both model fusion and data augmentation in improving speech recognition accuracy.

Table 2: The WER results of each ASR module.

Model	WER (%)
Baseline	13.53
Proposed system	10.25
w/o model fusion	11.95
w/o data augmentation	12.40

6.3.2 MT Module

We evaluate our MT module in terms of the BLEU score, which measures the n-gram overlap between the predicted output and the reference sentence.

Table 3: The BLEU results of each MT module.

Model	BLEU
Baseline	28.1
Proposed system	33.4
w/o three-stage fine-tuning	28.7
w/o back translation	30.8
w/o cross-validation	31.0

As shown in Table 4, the proposed system with three-stage fine-tuning achieves a significantly bet-

[7] https://github.com/huggingface/transformers

[8] https://github.com/jaywalnut310/vits

[9] https://github.com/kaldi-asr/kaldi

Table 4: Experimental results of TTS in terms of MOS and WER. BN means using two-stage training with BN and pre-trained spkr. embed. means using pre-trained speaker embedding.

Model	Clarity in CER (%)	Naturalness (MOS)	Sound Quality (MOS)	Speaker Similarity (MOS)
Baseline	7.14	3.38±0.05	3.81±0.04	2.12±0.06
Proposed system	6.12	3.70±0.06	3.86±0.06	3.72±0.06
w/o BN	7.12	3.40±0.04	3.81±0.05	3.10±0.07
w/o Pre-trained spkr. embd.	-	-	4.05±0.05	2.22±0.06
w/o Audio super-resolution	-	-	3.64±0.04	-
Recording	4.53	4.01±0.04	3.89±0.03	4.35±0.05

ter BLEU score than the baseline, demonstrating the effectiveness of curriculum learning in our scenario. Furthermore, by incorporating back translation and cross-validation, the translation performance can be further improved.

6.3.3 TTS Module

We calculate the character error rate (CER) to evaluate the clarity of speech for each TTS module. The ASR system used for calculating CER is the Chinese-English ASR model mentioned in Section 5.1.1. Additionally, we conduct mean opinion score (MOS) tests with ten listeners rating each sample on a scale of 1 (worst) to 5 (best) to evaluate naturalness, sound quality, and speaker similarity.

In the ablation study without pre-trained speaker embedding, speaker ID is to control the speaker timbre of the synthesized speech. To eliminate the influence of ASR and MT results on TTS evaluation, we use the Chinese text in the evaluation data and its corresponding English source speech as the reference of speaker timbre as the test set for TTS evaluation.

As shown in Table 3, our proposed system has achieved significant improvement in naturalness, sound quality, speaker similarity, and clarity of speech compared with the baseline. Interestingly, the system without pre-trained speaker embedding has better sound quality than both the proposed system and recording. We conjecture the reason is that the pre-trained speaker embedding greatly influences the sound quality in the zero-shot TTS setup. Therefore, the quality of the synthesized 24KHz audio is superior to the 16KHz recording, which can be demonstrated by the 3.64 MOS score of the system without audio super-resolution. Meanwhile, the speaker similarity MOS score is very low due to the lack of generalization ability to unseen speakers. Without using the BN-based two-stage model, the system decreases performance on all indicators, which shows the effectiveness of BN as an intermediate representation in our experimental scenario.

6.3.4 System Evaluation

Finally, we calculate the ASR-BLEU score for the baseline and the proposed system to evaluate the speech-to-speech translation performance. Specifically, we use the ASR system to transcribe the Chinese speech generated by TTS, and then compute the BLEU scores of the ASR-decoded text with respect to the reference English translations. The ASR system for transcribing Chinese speech is the same as that in Section 6.2.3.

Table 5: The ASR-BLEU results of each system.

Model	ASR-BLEU
Baseline	27.5
Proposed system	32.2

As shown in Table 5, our proposed system achieves a higher ASR-BLEU score than the baseline, which indicates that our proposed system has good speech-to-speech translation accuracy.

7 Conclusion

This paper describes the NPU-MSXF speech-to-speech translation system, which we develop for the IWSLT 2023 speech-to-speech translation task. Our system is built as a cascaded system that includes ASR, MT, and TTS modules. To ensure good performance with multi-source data, we improved each module using various techniques such as model fusion and data augmentation in the ASR, three-stage fine-tuning, back translation, and cross-validation in the MT, and two-stage training, pre-trained speaker embedding, and audio super-resolution in the TTS. Through extensive experiments, we demonstrate that our system achieves high translation accuracy, naturalness, sound quality, and speaker similarity with multi-source input.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondrej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussà, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 1–88. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 3670–3674. ISCA.

Yutian Chen, Yannis M. Assael, Brendan Shillingford, David Budden, Scott E. Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Çaglar Gülçehre, Aäron van den Oord, Oriol Vinyals, and Nando de Freitas. 2019. Sample efficient adaptive text-to-speech. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3830–3834. ISCA.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings*, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 1123–1127. ISCA.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J Han, and Shinji Watanabe. 2023. E-branchformer: Branchformer with enhanced merging for speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 84–91. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 3327–3339. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jae Soo Lim and Alan V Oppenheim. 1979. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Yanqing Liu, Zhihang Xu, Gang Wang, Kuan Chen, Bohan Li, Xu Tan, Jinzhu Li, Lei He, and Sheng Zhao. 2021. DelightfulTTS: The microsoft speech synthesis system for blizzard challenge 2021. *CoRR*, abs/2110.12612.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Masanori Morise. 2017. Harvest: A high-performance fundamental frequency estimator from speech signals. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, pages 2321–2325. ISCA.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, Jinsong Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Trans. Speech Audio Process.*, 14(2):365–376.

Markus Ojala and Gemma C. Garriga. 2010. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.*, 11:1833–1863.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 125–129. European Language Resources Association (ELRA).

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. AISHELL-3: A multi-speaker mandarin TTS corpus. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2756–2760. ISCA.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, 6(1):1–3.

Kun Song, Heyang Xue, Xinsheng Wang, Jian Cong, Yongmao Zhang, Lei Xie, Bing Yang, Xiong Zhang, and Dan Su. 2022a. AdaVITS: Tiny VITS for low computing resource speaker adaptation. In *13th International Symposium on Chinese Spoken Language Processing, ISCSLP 2022, Singapore, December 11-14, 2022*, pages 319–323. IEEE.

Kun Song, Yongmao Zhang, Yi Lei, Jian Cong, Hanzhao Li, Lei Xie, Gang He, and Jinfeng Bai. 2022b. DSPGAN: a gan-based universal vocoder for high-fidelity TTS by time-frequency domain supervision from DSP. *CoRR*, abs/2211.01087.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. 2022. GigaST: A 10, 000-hour pseudo speech translation corpus. *CoRR*, abs/2204.03939.

Yongmao Zhang, Heyang Xue, Hanzhao Li, Lei Xie, Tingwei Guo, Ruixiong Zhang, and Caixia Gong. 2022. Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. *CoRR*, abs/2211.02903.