The BIGAI Text-to-Speech Systems for Blizzard Challenge 2023

Zhihang Xie¹, Rui Fang¹, Mingtian Zhao¹

¹Beijing Institute of General Artificial Intelligence, China
zhihangxie@gmail.com

Abstract

This paper describes the text-to-speech systems submitted from Beijing Institute of General Artificial Intelligence to Blizzard Challenge 2023. Two French speech datasets are released in this year’s challenge, the Hub task (FH1) with a total duration around 51 hours and the Spoke task (FS1) with a total duration around 2 hours. The submitted systems are built upon the VITS model, which is a variational autoencoder for end-to-end speech synthesis. The base model is first trained in a self-supervised learning way to be robust to frontend errors and then finetuned on the target speech dataset for each task. The model inputs are phoneme sequences transcribed from French texts and represented as IPAs. The submitted systems are identified as L in the subjective evaluation and the listening results confirm that the synthetic speech is relatively natural and similar to the original speaker. The system ranks in the 3rd MOS score group in the Hub task, while the system ranks second in the Spoke task.

Index Terms: Blizzard Challenge, text-to-speech synthesis, speech adaptation, self-supervised learning

1. Introduction

Text-to-speech (TTS) or speech synthesis aims to generate more human-like speech waveforms from written texts. The Blizzard Challenge has been organized annually since 2005 [1, 2] in order to better understand and compare recent techniques in building TTS systems on the same data. Conventional speech synthesis methods like unit selection [3, 4, 5] and statistical parameter speech synthesis (SPSS) [6, 7, 8, 9] have been successfully applied in both academia and industry. Building a unit selection system often requires a large amount of manually labeled speech data, as well as expert knowledge to balance the weights among different features. Instead, SPSS systems have the advantages of small footprint, robustness and flexibility to change voice characteristics, but the quality of synthetic speech underperforms that of unit selection systems due to vocoding, accuracy of acoustic models and oversmoothing.

These problems have been well studied in recent years and the speech quality has been greatly improved with the advancements of deeper and larger neural networks. Apart from text processing modules, TTS system pipelines are simplified to two-stage generative modeling. The first stage applies an acoustic model to generate intermediate speech representations. End-to-end acoustic models like Tacotron 1/2 [10, 11] and FastSpeech 1/2 [12, 13] are proposed to generate acoustic features from a sequence of phonemes rather than a sequence of linguistic features. The second stage applies a vocoder to generate raw waveforms conditioned on the intermediate representations. More neural vocoders are proposed to replace traditional vocoders [14, 15] with autoregressive vocoders [16, 17], flow-based vocoders [18, 19], GAN-based vocoders [20, 21], diffusion-based vocoders [22, 23, 24], etc. Models at each stage are trained separately and assembled during inference.

Further investigations [25, 26] have proved the effectiveness of single-stage end-to-end TTS systems, which can avoid error propagation in two-stage systems and reduce the efforts for manual annotations and feature engineering. In this work, the VITS¹ model [26] is employed as the backbone architecture to generate synthetic speech for the two tasks. The training process is divided into three stages: 1) train a pBART model on French phonemes in a self-supervised learning way; 2) train a base VITS model on the Common Voice dataset; 3) finetune the base model on the target dataset for each task.

The rest of this paper is organized as follows. Section 2 describes which datasets are utilized in the training phase. Section 3 explains how to process input texts and how to train the models. Section 4 shows subjective evaluation results by the organizers. Section 5 concludes the submissions.

2. Data

There are three speech datasets used to train the models and statistics on these datasets are shown in Table 1.

2.1. Mozilla Common Voice

The Mozilla’s Common Voice project [27] provides a multilingual collection of transcribed speech intended mostly for speech recognition applications. The French V12.0 dataset is expected to be a good resource to pretrain the base model. To keep consistent with the gender of target speakers in the two tasks, utterances explicitly labeled with the gender of female are selected. All the audio files are unified to single-channel 22050Hz format. The dataset is comprised of 57,068 utterances recorded by 598 speakers with the total duration around 84.23 hours.

2.2. Blizzard Challenge 2023

There are two tasks in this year’s challenge, the Hub task for speech synthesis and the Spoke task for speaker adaptation.

Table 1: Statistics on speech datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Speakers</th>
<th>Utterances</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hub Task</td>
<td>1</td>
<td>63,689</td>
<td>51.26</td>
</tr>
<tr>
<td>Spoke Task</td>
<td>1</td>
<td>2,494</td>
<td>2.05</td>
</tr>
<tr>
<td>Common Voice</td>
<td>598</td>
<td>57,068</td>
<td>84.23</td>
</tr>
</tbody>
</table>

¹The original implementation: https://github.com/jaywalnut310/vits
2.2.1. Hub Task

The Hub task contains 63,689 utterances recorded by a female speaker with the total duration around 51.26 hours, where sentences are extracted from 289 chapters in audiobooks from LibriVox written by 5 authors. It is observed that utterances in this dataset is much shorter than those in the Common Voice dataset.

2.2.2. Spoke Task

The Spoke task contains 2,494 utterances recorded by another female speaker with the total duration around 2.05 hours, where sentences are selected from various books and transcripts of French parliament sessions.

3. Method

3.1. Text Processing

Conventional speech synthesis systems [28, 29, 30] transform orthographic representations into a sequence of linguistic specifications using a text analyzer like Festival [28], which is a pipeline system mainly composed of modules like text normalization, homograph disambiguation, grapheme-to-phoneme conversion and prosody prediction. However, building these modules often requires expert knowledge to construct lexicons and rules, as well as many user-friendly functions is chosen to simplify this process.

Private annotations are used instead of the provided annotations to keep consistency between the training data and the testing data. To be specific, the Phonemizer backend is set to espeak with phonemes represented using international phonetic alphabets (IPAs). Without linguistic knowledge of French, the phone set is not further split into different categories as provided by the organizers. According to the instructions, punctuation and special symbols are preserved to indicate prosody clues like speaking quotes, turn switches, three dots, quoted expletions and special symbols are preserved to indicate prosody.

3.2. Self-supervised Learning

Following the BART pretraining scheme [32], a phoneme BART (pBART) model consists of the text encoder of the VITS model and a 6-layer Transformer decoder. Given a set of phoneme sequences $D$, the model learns to reconstruct corrupted phoneme sentences $g(x)$ to their original forms $x$. The pretraining objective $\mathcal{L}_d$ defined as Equation 1 is a denoising loss, where $g$ is a denoising function that corrupts the original sequence, and $P$ is defined by the model. As described in [32], there are two types of noise in $g$, random masking and random substitution. Within the 35% randomly selected phonemes in $x$, 80% are replaced by the mask token, 10% are replaced by a random phoneme from $x$ and the rest 10% remain unchanged.

$$\mathcal{L}_d = \sum_{x \in D} \log P(x|g(x); \theta)$$  \hspace{1cm} (1)

The pBART decoder has 6 Transformer layers with 4 attention heads, model dimension of 256 and inner dimension of 256. The total 18.4M parameters are optimized by the Adam optimizer [34] with the initial learning rate and annealing factor set to $10^{-4}$ and 0.9 respectively. The learning rates are updated based on the improvement of the training losses between the previous epoch and the current epoch. The 123,251 transcribed phoneme sequences are split into training and validation subsets with the ratios of 0.9 and 0.1. The batch size is set to 6 and the model is trained for about 150 epochs.

3.3. Finetuning

The base model has exactly the same architecture as the VITS model as shown in Figure 1, which consists of a posterior encoder, prior encoder, a decoder, a discriminator and a stochastic duration predictor. The posterior encoder and discriminator are only used for training. The pBART encoder is used to initialize the phoneme encoder, while the pBART decoder is discarded in the finetuning stages. Given the total number of speakers, speaker embedding will be produced and passed through the posterior encoder, stochastic duration predictor and decoder. During finetuning and inference, the target speaker will be specified to distinguish the two tasks.

The base model is first trained on the Common Voice dataset. The total 98.1M parameters are optimized by the AdamW optimizer [34] with the initial learning rate and annealing factor set to $2e^{-4}$ and $0.999^{1/8}$ respectively. The model is set to 64 and is trained for 300k steps. Then, the base model is finetuned on the Hub task and the Spoke task respectively.

4. Results

4.1. Hub task

In this task, 21 systems are evaluated in total with 18 from participating teams denoted as C to T, 2 from the organizers denoted as BF for FastSpeech2 [13] + HiFiGAN and BT for Tacotron2 [11] + HiFiGAN [21] and 1 from natural speech denoted as A. The subjective evaluation contains 5 sections, 2 naturalness tests, 1 similarity test and 2 intelligibility tests (1 for semantically unpredictable sentences and 1 for homographs).

4.1.1. Naturalness

In this section, listeners listened to one audio sample at a time and chose to score the quality of the audio on a scale from 1 (very poor) to 5 (excellent). The total 361 listeners consist of 312 paid listeners, 39 speech experts and 10 online volunteers. Figure 2 shows the boxplot of naturalness scores for all listeners. A hierarchical clustering of the systems is represented as Figure 3 so that systems with similar scores are grouped into small clusters. The submitted system denoted as L is classified into the 3rd group and ranks 16th out of 21.
4.1.2. Similarity
In this section, listeners compared one given audio sample at a time with four reference samples of the original speaker and chose to score the similarity between them on a scale from 1 (completely different person) to 5 (Exactly the same person). The total 348 listeners consist of 305 paid listeners, 37 speech experts and 6 online volunteers. Figure 4 shows the boxplot of similarity scores for all listeners. The submitted system L is classified into the 3rd group and ranks 13th out of 21. It is noticed that system F and system M perform even better than natural speech A by the original speaker.

4.1.3. Intelligibility
In this section, listeners listened to one audio sample at a time and typed in what they heard according to the spelling rules of French. The sentences are designed to remove semantically meanings to avoid ceiling effects and variants of some French words are accepted. All the 228 listeners are paid listeners and native speakers. Figure 5 shows word error rates (WERs) on the sentences and the submitted system L ranks 2nd to the last.

4.1.4. Homograph
In this section, listeners first listened to three audio samples, two of which are uttered by two reference speakers. Then they were asked to select the reference audio that corresponded the best to the pronunciation of the homograph, regardless of the correctness of the pronunciation. All the 218 listeners are either paid listeners or speech experts. Figure 6 shows the pronunciation error rates (PERs) of homographs in the sentences and the submitted system ranks 14th out of 20.

4.2. Spoke task
In this task, 17 systems are evaluated in total with 14 from participating teams, 2 from the organizers and 1 from natural speech. The subjective evaluation contains 3 sections, 2 naturalness tests and 1 similarity test.

4.2.1. Naturalness
In this section, the total 282 listeners consist of 245 paid listeners, 30 speech experts and 7 online volunteers. Figure 7 shows the boxplot of naturalness scores for all listeners and the hierarchical clustering plot is shown in Figure 8. The submitted system denoted as L is classified into the 2nd group and ranks 4th out of 17. It is found that system F is the only one with a higher MOS score than natural speech A.

4.2.2. MUSHRA
In this section, listeners listened to one explicit reference audio sample and 6 non-identified audio samples, one hidden refer-
ence, one BF reference and 4 systems with the highest MOS quality scores. They were asked to score the non-identified audio samples on a scale from 0 (very poor) to 100 (excellent). The scores are first sorted in order from the worst to the best and then refined as many times as the listeners like. The total 47 listeners consist of 28 paid listeners, 18 speech experts and 1 online volunteers. Figure 9 shows the boxplot of MUSHRA scores for all listeners. It is surprising to find that the ranking of the submitted system L has risen from 4th to 3rd, after natural speech A at 1st and system F at 2nd respectively.

Further investigations confirm that the differences among natural speech A, system F and system L are significant, while system L and system O are not significantly different. Figure 10 shows the binary significant differences between the six systems. Moreover, the organizers applied post-hoc multiple comparisons between the levels of the system factor [35], where model statistics are 33.755 between A and F, 9.4541 between F and L and 0.7295 between L and O.

4.2.3. Similarity

In this section, the total 286 listeners consist of 243 paid listeners, 31 speech experts and 12 online volunteers. Figure 11 shows the boxplot of similarity scores for all listeners. The submitted system L is classified into the 1st group and ranks 4th out of 17. Noticeably, the number of systems with higher scores than natural speech A increases to five in total.

5. Conclusion

This paper presents the BIGAI systems submitted to Blizzard Challenge 2023. The systems first transcribe French texts into phonemes and then generate speech waveform directly using the VITS models. Some methods are applied to improve the performance like self-supervised learning and multispeaker pre-training. Subjective evaluation results confirm that 1) text processing is still important especially for homographs; 2) the quality of synthetic speech is relatively natural and similar to the original speaker; 3) the proposed method works remarkably well given very limited data.
6. References


