**Abstract**

This paper presents the 10AI text-to-speech (TTS) system "Xpress" for the Blizzard Challenge 2023. Xpress is an end-to-end TTS system trained on the released publicly available French speech data for Hub task 2023-FH1. The Xpress system is constructed with a character-level prosody modeling component, a conditional Flow Variational AutoEncoder (FV AE) based acoustic model and a BigVGAN vocoder. Comparing with conventional TTS systems, no conventional front-end analysis component is employed in the proposed Xpress, thus significantly simplifying the overall construction of the TTS system. Moreover, with the simplified FV AE acoustic model design, the Xpress requires less computation resource in training and inference, donates three times faster in inference than transformer-style system. Evaluation results demonstrate the proposed system with competitive effectiveness in both MOS quality and MUSHRA evaluation, and leading quality in both SUS and HOMOS intelligibility assessments.

**Index Terms:** text-to-speech synthesis, end-to-end, FVAE.

1. Introduction

The Blizzard Challenge, organized annually since 2005 [1], aims to advance text-to-speech (TTS) technologies by comparing and analyzing various approaches. Participants are required to extract the released available data and construct synthetic voices. The performance of synthetic voices is evaluated through subjective listening tests and objective metrics. For this year, the challenge is designed for French speech synthesis and contains two tasks: (1) a Hub task to build a French voice using only the provided publicly available data; (2) a Spoke task to construct a target French voice with no training data limitation. In this challenge, we proposed our latest Xpress TTS system and enrolled in the Hub task 2023-FH1 to be compared with different state-of-the-art (SOTA) approaches.

Text-to-speech synthesis, aiming to build natural and expressive human-like speech, remains a challenging and fascinating research field despite decades of development [2]. A general TTS system utilizes several components to form a complex pipeline [3] for speech synthesis, including front-end text analysis, back-end acoustic modeling and vocoder. With recent development of deep learning and neural networks, a list of high quality text-to-speech systems have been developed and proposed. For conventional approaches, such as rich unit based concatenative synthesis [4] and statistical parametric speech synthesis (SPSS) [5] [6], integrating deep neural networks into the data preparation pipeline and different TTS components can significantly enhance their synthesis quality and naturalness. However, the front-end and back-end TTS components in these conventional systems are trained separately, making the process vulnerable to errors compounding through the different stages [7, 8]. End-to-end TTS systems therefore are developed to address this shortcoming [8] [9]. Compared to conventional TTS systems, the novel end-to-end TTS systems integrate TTS components into a single trainable framework and provide more natural and expressive speech synthesis performance.

Tacotron [8] introduced the direct mapping from raw character sequence to acoustic features by using an attention-based autoregressive acoustic modeling. Furthermore, Tacotron 2 [9] integrates WaveNet [10] vocoder into the trainable structure for higher generation quality. These systems have achieved improvements in synthesis quality and naturalness improvement compared with the conventional TTS approaches, but still have obvious limitation in training efficiency and inference robustness for the autoregressive design. Transformer based TTS system [11] is therefore proposed to improve the efficiency of end-to-end TTS systems in training on large data, and the Robust-Tacotron [12] is proposed for robustness enhancement in inference. These approaches have effectively solved the weakness of end-to-end TTS systems, but are still limited in inference efficiency for the autoregressive nature.

FastSpeech [13] is firstly proposed to address the issue in autoregressive TTS systems. Instead of using embedded speech duration constraints, FastSpeech employs a trainable duration predictor and feed-forward transformers in encoder and decoder for parallel acoustic feature prediction. It overcomes the autoregressive limitations and achieves much better efficiency and robustness compared with autoregressive approaches. FastSpeech 2 [14] further improves the accuracy of the duration predictor and introduces more speech variation information, such as pitch and energy as conditional inputs, to enhance the diversity of synthesized speeches. DelightfulTTS [15] is then proposed to further model variation information from both implicit and explicit perspectives and achieve SOTA quality in Blizzard Challenge 2021. In general, the primary objectives of modern TTS systems are as follows [16]: fast, lightweight, high-quality, expressive and diverse.

Following the primary developing goals, we design and bring the novel conditional Flow Variational AutoEncoder (FVAE) based TTS system “Xpress” to the Blizzard Challenge 2023. The Xpress system is constructed with a FVAE based acoustic model and a BigVGAN vocoder. The characteristics of the Xpress TTS system could be summarized as:

- Comparing with conventional TTS systems, none conventional front-end analysis component is employed in the proposed Xpress, thus significantly simplifying the overall construction of the TTS system, the intelligibility assessments have stated the effectiveness of this structure.
- With well-designed flow based acoustic modeling structure, Xpress can produce synthetic speech with high quality and...
naturalness, being on par with other SOTA TTS systems in MOS quality and MUSHRA assessments.

- The Xpress is designed to be a lightweight TTS system, costing less computational resource in training and inference, and being three times faster than FastSpeech in inference.

The following paper is organized as follows: Section 2 describes the architecture of the proposed system in detail. Section 3 provides a comprehensive discussion of the evaluation results in the Blizzard Challenge 2023. Section 4 presents the conclusion and future work.

2. Proposed Xpress System

The overall structure of Xpress is shown in figure 1. Inspired by and following PortaSpeech [16], Xpress is constructed with two main components: a conditional Flow Variational AutoEncoder (FVAE) based acoustic model and a BigVGAN vocoder [17]. In modeling, the FVAE encoder firstly encodes the acoustic features to a sequence of character-level prosody latent variable $z$. Then, the FVAE decoder decodes the latent back to acoustic features. In inference, latent variable $z$ is predicted using the linguistic features instead of using a random value to enhance the prosody of generated speech.

2.1. Flow Modeling

Conventional frame-level prosody features and word-level prosody features can be harmful to the prosody reconstruction in TTS systems, since the frame-level prosody features can cause significant linguistic information leakage and the word-level prosody features are too coarse to capture the finer details of prosody [18]. The character-level prosody features are thus proposed for use in Xpress. To align the character-level prosody features and targeted frame-level acoustic features, a character-level duration prediction component is employed to construct a hard alignment matrix. This matrix represents the correspondence between character inputs and the targeted acoustic frames, allowing Xpress to convert the frame-level prosody features to the character-level prosody features in training.

$$PF_{char} = A \cdot PF_{frame}$$

Where the variables $PF_{char} \in \mathbb{R}^{T_s \times d}$ and $PF_{frame} \in \mathbb{R}^{T_f \times d}$ represent the character-level and frame-level prosody features. The hard alignment matrix $A \in \mathbb{R}^{T_s \times T_f}$ is used to denote the correspondence between characters and acoustic frames and is learned from the training data.

The processes inside the FVAE encoder follow two stages: In the first stage, the input acoustic features, with the acoustic-information-expanded linguistic features as a condition, are sent into the encoder for hidden states computing. Subsequently, the alignment matrix $A$ is utilized to downsample the hidden states to character-level. In the second stage, the encoder processes the character-level hidden states to obtain the latent variables $z$ for further modeling. The process inside FVAE decoder is totally inverted to FVAE encoder, the extracted character-level latent is upsampled to frame-level by using the inverse of alignment matrix $A$ and are then mapped to original acoustic features. The upsampling module is constructed following the proposed [14], to produce frame-level hidden states with duplication of the character embeddings.

Specifically, comparing with conventional Variational AutoEncoder (VAE), the proposed system employs a small volume-preserving normalizing flow [16] as res flow block, to transform simple distributions, such as the Gaussian distribution, into complex distributions through a series of $K$ invertible mappings. This design can unfreeze the posterior constraints limitation in conventional Gaussian distribution [19], providing stronger diversity and generative capacity in modeling. These mappings are implemented as a stack of convolutional neural network residual blocks with a dilatation at 1.

2.2. Prosody Predictor

Inspired by DelightfulTTS [15], a prosody predictor is employed to model the correspondence between linguistic input and character-level prosody. In modeling, the prosody predictor is designed to map the latent variables $z$ by minimizing the KL divergence between the predicted normal distribution and the posterior distribution, with the encoded result from text encoder as input. In inference, the well-trained prosody predictor can directly predict the character-level prosody features from given text input. Specifically, in practice, the character-level prosody is difficult to model from text directly, thus additional character-level pitch and energy information from audio is used for modeling as proposed in [14].

2.3. Vocoder Construction

To minimize the modeling complexity of acoustic model, mel spectrogram extracted from 16kHz waveform is selected as the acoustic features in modeling [15]. BigVGAN [17] with residual connection presented in FreGAN [20] is employed to convert the generated 16kHz mel spectrogram into 22.05kHz waveform. During the training stage, the time and frequency do-
mains integrated discriminator from UnivNet [21] is employed, which is to provide adversarial training loss, to optimize overall training cost and improve speech generation quality.

2.4. Data Processing

The training data provided by the organizers comprises approximately 50 hours of speech data from a female native French speaker. In the Hub task, no additional data is allowed to be used in training. During processing, the Montreal Forced Aligner [22] is firstly employed to align the text and audio, tagging character-level timestamp on each individual character. Subsequently, two to five short sentence segments are randomly selected and concatenated to form longer sentences for data augmentation purposes. Finally, all audios are sent to resample to 16kHz, trim the silence, and get the loudness be normalized. Then 80-dimensional Mel spectrums are extracted using 40ms frame length and 10ms frame hop with Hann window function enabled.

2.5. Model Training

In training, 100 speech sentences from provided dataset are employed as validation set, and rest sentences are used as training set. The proposed Xpress is trained with Adam [23] optimizer by 300k steps, started with a learning rate at 0.001, and a dynamic batch size contains maximum 48000 frames for each step. Specially, BigVGAN vocoder is trained from scratch, with no additional speech data employed.

2.6. Inference Efficiency

The transformer-based linguistic encoder is constructed 4 transformer layers, each contains 512 units, the prosody predictor is constructed with 6 convolutional layers, each contains 512 units. The FVAE-based encoder and decoder is constructed following the proposed in [16]. The inference efficiency of Xpress is evaluated by comparing it with FastSpeech in the same hardware environment using the same testing scripts. Specifically, before testing, an ABX test is processed, and the test result shows the generation quality of these two systems is on par. The performance evaluation result is shown in the Table 1.

Table 1: The comparison result of inference efficiency. The evaluation is conducted on a server with AMD EPYC 7K83 CPU, only one core is employed for inference, and the text set is the same.

<table>
<thead>
<tr>
<th>Method</th>
<th>RTF</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastSpeech</td>
<td>0.2</td>
<td>40M</td>
</tr>
<tr>
<td>Xpress</td>
<td>0.05</td>
<td>15M</td>
</tr>
</tbody>
</table>

3. Evaluation Results

The listening evaluation of Blizzard Challenge 2023 contains five different tests, including quality and similarity MOS (Mean Opinion Score) assessments, MUSHRA (Multi Stimulus test with Hidden Reference and Anchor) quality assessment, SUS (Semantically Unpredictable Sentences) and HOMOS (Homographs of Sentences) intelligibility assessments. For MOS quality assessment, for each panel, judges were asked to listen to one audio sample and score it from 1 to 5 on the audio quality. For MOS similarity, judges would listen to four reference audios for each of the 7 stimuli and were asked to score the similarity between given synthesis audio and reference on a 5-scale score. An additional MUSHRA is employed to refine quality measurements of the best three systems in MOS quality assessments. For SUS intelligibility assessment, judges are required to transcribe words from given synthetic speech, and the word error rate (WER) is then calculated. For the HOMOS intelligibility assessment, judges are required to select the closest homographs in the given synthetic speech and two references homographs audios. Hundreds of judges have participated in the evaluations, and thanks to their incredible work, we obtained significant and meaningful results for the challenge. In the evaluation, natural speech is identified as A, System BF is a benchmark system based on FastSpeech2 with HiFi-GAN vocoder, and System BT is a benchmark system based on Tacotron2 with HiFi-GAN vocoder. Systems C to T are the challenge participants.

3.1. Quality MOS

The quality MOS result is shown in figure 2. In this test, 361 judges were involved, and the proposed system (mean score 4.3) achieved a subjective perception very close to natural speech (system A, mean score 4.4). This demonstrates the effectiveness of the proposed system.

3.2. Similarity MOS

Similarity MOS result is shown in figure 3. In this test, 348 judges have enrolled and the proposed system has achieved a limited score. This result is contrary to our previous internal assessment results, in which the proposed system has achieved very close score towards natural speech. However, since the given training data is constructed from audio books, the speaker prosody or even timbre diversity is quite rich. It could be hard for the proposed system to produce speech with accurate expression until further context information is provided.

3.3. MUSHRA

MUSHRA result is shown in figure 4. In this test, 47 native judges including 18 language experts have enrolled, and the proposed system has achieved a close subjective perception towards other systems, and shows its superiority over baseline system. This result further demonstrates that the proposed system is competitive with other SOTA systems.
3.4. SUS Intelligibility

SUS intelligibility assessment result is shown in figure 5. In this test, 228 judges have enrolled and the proposed system has achieved the WER on par with other SOTA systems. This result demonstrates the proposed front-end-free system can fully finish the linguistic analysis tasks implicitly and achieve intelligibility quality on par with SOTA systems.

3.5. HOMOS Intelligibility

HOMOS intelligibility assessment result is shown in figure 6. In this test, 218 judges have enrolled and the proposed system has achieved acceptable pronunciation accuracy. This result demonstrates the proposed front-end-free system can handle homographs, however, still have significant gap towards SOTA systems with well-designed homographs analysis components.

3.6. Discussion

The proposed Xpress is an end-to-end TTS system, and none conventional front-end analysis component is employed for word segmentation, Part-of-Speech tagging, break prediction, etc. In its construction, the system only employs a simple FVAE encoder-decoder structure with additional character prosody predictor to map the input characters towards acoustic features. This simple but powerful structure supports the proposed Xpress system to generate natural speech on par with SOTA systems in both quality and intelligibility with better computing efficiency.

4. Conclusions

This paper presents the 10AI TTS synthesis system for Blizzard Challenge 2023, a concise end-to-end system using characters as input. The proposed system has achieved performance on par with SOTA systems in MOS quality and MUSHRA assessments, while being one of the leading systems in SUS and HOMOS intelligibility assessments. These results demonstrate the effectiveness of the proposed system, indicating the possibility to make TTS systems simpler, smaller, and faster.

Further work. The Xpress TTS system is still under developing, the overall quality and naturalness of Xpress are still limited by its ability to synthesize expressions for different emotions, scenarios, and diversities. We are still working on integrating the latest research findings into this system, to bring spontaneous inspirations to the TTS systems for better and expressive speech generations.

5. Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities No. 2023RC03.
6. References


