# Accented Text-to-Speech Synthesis With Limited Data

Xuehao Zhou, *Student Member, IEEE*, Mingyang Zhang, *Member, IEEE*, Yi Zhou, *Member, IEEE*, Zhizheng Wu, *Senior Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—This paper presents an accented text-to-speech (TTS) synthesis framework with limited training data. We study two aspects concerning accent rendering: phonetic (phoneme difference) and prosodic (pitch pattern and phoneme duration) variations. The proposed accented TTS framework consists of two models: an accented front-end for grapheme-to-phoneme (G2P) conversion and an accented acoustic model with integrated pitch and duration predictors for phoneme-to-Mel-spectrogram prediction. The accented front-end directly models the phonetic variation, while the accented acoustic model explicitly controls the prosodic variation. Specifically, both models are first pre-trained on a large amount of data, then only the accent-related layers are fine-tuned on a limited amount of data for the target accent. In the experiments, speech data of three English accents, i.e., General American English, Irish English, and British English Received Pronunciation, are used for pre-training. The pre-trained models are then fine-tuned with Scottish and General Australian English accents, respectively. Both objective and subjective evaluation results show that the accented TTS front-end fine-tuned with a small accented phonetic lexicon (5 $k$ words) effectively handles the phonetic variation of accents, while the accented TTS acoustic model fine-tuned with a limited amount of accented speech data (approximately 3 minutes) effectively improves the prosodic rendering including pitch and duration. The overall accent modeling contributes to improved speech quality and accent similarity.**

***Index Terms*—Text-to-speech (TTS), accent, phonetic variation, prosodic variation.**

Manuscript received 4 June 2023; revised 28 December 2023 and 16 January 2024; accepted 23 January 2024. Date of publication 16 February 2024; date of current version 6 March 2024. The work of Haizhou Li was supported in part by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen under Grant B10120210117-KP02, in part by the Shenzhen Natural Science Foundation Key Project under Grant JCYJ20220818103001002, and in part by the National Natural Science Foundation of China under Grant 62271432. The work of Zhizheng Wu is supported by the National Natural Science Foundation of China under Grant 62376237. This work was supported in part by the Agency for Science, Technology and Research (A*STAR) Human Robot Collaborative AI through AME Programmatic Funding Scheme under Project A18A2b0046. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kai Yu. *(Corresponding author: Zhizheng Wu.)*

Xuehao Zhou and Yi Zhou are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: xuehao.zhou@u.nus.edu; yi.zhou@u.nus.edu).

Mingyang Zhang and Zhizheng Wu are with the Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: zhangmingyang@cuhk.edu.cn; wuzhizheng@cuhk.edu.cn).

Haizhou Li is with the Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China, also with Kriston AI, Xiamen 361000, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077 (e-mail: haizhouli@cuhk.edu.cn).

Digital Object Identifier 10.1109/TASLP.2024.3363414

## I. Introduction

TEXT-TO-SPEECH (TTS) synthesis aims to synthesize high-quality speech from the given text input. Traditional TTS pipelines are based on statistical parametric modeling [1], [2], [3], which involves multiple components in the training phase. Recently, end-to-end (E2E) neural TTS systems [4], [5], [6], [7], [8] achieve state-of-the-art performance, generating human-like natural speech directly from text input. However, to meet different customer requirements in real-world applications, accented TTS is required for an improved user experience. Accented TTS can help people with different accents or dialects to express themselves authentically and also allow them to use more natural and familiar sounds, improving their overall communication experience. Another potential use case for accented TTS is the customization of generated voices to users' accent preferences. This personalized service can increase user engagement, satisfaction and adoption with customer service, voice assistants and audiobooks. However, transcribing a large amount of accented speech requires a lot of effort, and in practice, extensive recordings of accented speech from low-resource languages are not always available. The problem of data scarcity in accented speech makes it necessary to create an accented TTS with limited data.

Unlike multi-speaker TTS [9], [10], [11] systems that produce the voice of a target speaker by conditioning on an utterance-level speaker representation, or emotional TTS [12], [13], [14] systems that represent the specific emotion with an utterance-level emotion embedding, an accent is characterized at different levels of an utterance [15]. Therefore, it is challenging to preserve the speaker identity and accent attributes simultaneously in the generated speech from TTS, especially when the available data for the target accent is limited. This study is motivated to address such problems in accented TTS synthesis.

A foreign accent occurs when a native speaker of the first language (L1) pronounces the second language (L2) due to different linguistic systems between L1 and L2 [16], [17]. The pronunciation pattern of the segmental and suprasegmental structures affects the perception of a foreign accent [18], [19]. The attributes of a foreign accent can be mainly categorized into the variations of phoneme and prosody [19], both of which are essential components of accent representation [20], [21]. Taking English as an example, the phonetic transcriptions of lexical words vary from accent to accent. Such differences can be described in two aspects [22]: (a) the difference between the

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

phoneme sets, e.g., the vowel set /ax, ea, ia, ua, oh/ does not appear in both British and American English accents; (b) the difference between the phonetic transcriptions, e.g., the word 'day' is transcribed as /dei/ in the American accent, while /dæi/ in the Australian accent. On the other hand, prosody attributes such as pitch [23] and duration are shown to be effective in accent morphing [22]. In a study on Turkish [24], results suggest that fundamental frequency ($F0$) and duration play a critical role in the rendering of accents. Obvious differences are observed between lexically accented and unaccented syllables with respect to $F0$ peaks. In [25], [26], $F0$ and duration are selectively transplanted to observe that the two prosodic cues significantly affect accentedness.

To build an accented TTS system, the research problem is how to effectively model the accent-specific phonetic and prosodic patterns. In this work, we consider that an accent can be characterized by three key attributes, namely, phonetic variation, pitch pattern, and phoneme duration. We propose a framework for accented TTS that consists of an accented front-end for grapheme-to-phoneme (G2P) conversion and an accented acoustic model for phoneme-to-Mel-spectrogram prediction. We then study ways to build an accented TTS system with a limited amount of training data. In particular, we leverage the pre-trained models to transfer the shared knowledge from the large-scale general data to an accented TTS system.

In this paper, we propose an accented front-end to model the accent-specific phonetic variation. The front-end [27] learns to convert input text into a phonetic sequence in the target accent. Meanwhile, we also integrate a pitch predictor and a duration predictor into the acoustic model to modulate the $F0$ and phoneme duration for accent rendering. In order to investigate the respective effectiveness of lexicon and speech data on generating accented speech under the condition of limited data, the proposed techniques are validated in two scenarios. In the first scenario, we assume that we have only a small accented phonetic lexicon and explore how to handle the phonetic variation in the TTS front-end. In the second scenario, we further assume that we not only have a small accented phonetic lexicon, but also a small amount of accented speech data. With both limited lexicon and speech data, we explore how prosodic variation can be modeled alongside the phonetic variation. The main contributions of this work are summarized below:

- We formulate the accent rendering problem by addressing three key attributes of speech, namely, phonetic variation, pitch pattern, and phoneme duration.
- We address the phonetic variation problem by fine-tuning the TTS front-end with a small accented phonetic lexicon.
- We control the prosody of speech by incorporating $F0$ and duration predictors in the TTS acoustic model, which can be fine-tuned with limited accented speech data. The $F0$ and duration are further adopted as additional inputs to the attention-based decoder.

The rest of the paper is organized as follows. In Section II, we introduce the related work to set the stage of this study. Our proposed method and framework are presented in Section III. The experimental setup and result analysis are shown in Section IV and V, respectively. Section VI concludes this paper.

## II. Related Work

### A. Traditional Accented TTS Synthesis

Accented speech synthesis for various languages has been investigated. In [28], [29], Indian accented speech synthesis is studied with Festvox, a unit selection tool. Kolluru el at. [30] propose a method to generate multi-accent pronunciations for individual speaker by building a space of accents. They convert the phoneme sequence in the canonical version to an accented sequence via joint sequence model interpolation. The interpolated weights as points specify the different accents within an accent space. This method effectively changes the pronunciation of phonemes across accents. However, the prosodic variation, e.g., pitch and phoneme duration, of accent attributes is not modeled explicitly.

Anumanchipalli et al. [31] present an intonation model that automatically predicts the appropriate intonation contours from text for a statistical parametric speech synthesis system. This intonation model takes the accent group, a sequence of intonation events, as the modeling unit. This work reveals that the unit of accent components, such as $F0$, has an extremely high correlation with the linguistic pattern from text, which motivates us to consider using $F0$ to control the prosodic variation in accent rendering.

### B. Neural E2E Accented TTS Synthesis

With the advanced neural-network-based approaches, Abeysinghe et al. [32] adopt the E2E TTS architecture for accented speech generation. They pre-train a native TTS model with an American speech corpus and then perform fine-tuning with a non-native speech corpus. They visualize and analyze the vowel space variation during the fine-tuning stage and claim that the vowel space of voice generated from the fine-tuned model is closer to the non-native speech than to the native speech. Liu et al. [33] present a controllable accented TTS framework. They propose an accent intensity modeling method that quantifies the accent intensity for each sample and a consistency constraint loss subject to the total TTS training loss. Their results demonstrate the effectiveness of using intensity control on accent rendering. Melechovsky et al. [34] propose a controllable speech synthesis system based on a conditional variational autoencoder. Their proposed method is capable of generating a specific speaker's voice with an arbitrary target accent.

However, none of the above works address the phonetic variation of English accents, which is a significant aspect in accent morphing. Moreover, prior studies typically rely on a large amount of training data, which limits the scope of applications with low-resource settings.

In this paper, we seek to model both phonetic and prosodic variations in accented TTS with limited training data. To model the prosodic information, Yasuda et al. [35] develop an E2E Japanese speech synthesis system by capturing long-term dependencies related to pitch accents from text encoder with an additional self-attention layer. In addition, Yasuda et al. [36] examine that the pre-trained PnG BERT can capture the


Fig. 1. System diagram of a typical TTS synthesis pipeline. Mel refers to the acoustic feature Mel-spectrogram, while Wav is the speech waveform.

information related to pitch accents for a Japanese TTS system. They perform fine-tuning PnG BERT together with a TTS system and a tone prediction task to force PnG BERT to enrich the pitch accent information. Therefore, we consider controlling prosodic information of accent rendering from the text encoder and extending it to accented speech generation in limited data scenarios.

## III. Methodology

A TTS system is generally a pipeline of three components, as shown in Fig. 1: front-end, acoustic model, and vocoder. The front-end plays a crucial role in a TTS system to provide the required phoneme-relevant linguistic knowledge [37], [38], [39]. We propose an accented TTS system that consists of an accented front-end and an accented acoustic model. The accented front-end model transcribes the text input into an accented phoneme representation, that handles the accented phonetic variation. We use one of the state-of-the-art E2E TTS systems, Tacotron 2 [5] architecture, as the backbone of our accented acoustic model to predict the Mel-spectrogram with rich accented representations. The Parallel WaveGAN [40] is adopted as the neural vocoder to generate the speech waveform in the time-domain from the predicted Mel-spectrogram.

We study two scenarios on the accented TTS framework, a) only a small accented phonetic lexicon is available for the target accent. b) both a small accented phonetic lexicon and limited accented speech samples are available for the target accent.

### A. Accented TTS Framework

*1) Accented Front-End:* The Transformer-based model [41] is well-known and effective for converting graphemes into phonemes. However, it requires a large phonetic lexicon to build a standard Transformer-based G2P model, which is unrealistic for various English accents. It has been reported [42], [43] that the performance of low-resource G2P models are improved by transferring shared knowledge from pre-trained models. Thus, we employ the technique of pre-training followed by fine-tuning to build an accented front-end with a limited size of accented phonetic lexicon.

Inspired by the studies in multi-lingual and multi-speaker speech synthesis [44], [45], [46], we propose to pre-train a multi-accent G2P model with multiple English phonetic lexicons and fine-tune the accent-related layers with a small accented phonetic lexicon for the target accent. A multi-accent G2P approach is motivated by the idea of the multi-lingual G2P model, where a G2P system is built to model multiple phonetic systems. The former deals with accents, while the latter deals with languages. An early study [47] takes the language identity (ID) as an additional input for training a multi-lingual G2P model. We use an accent ID instead to pre-train a multi-accent G2P model in this work.

The left panel of Fig. 2 is the multi-accent front-end for G2P conversion. Overall, the G2P model takes both an English grapheme sequence and an accent ID as the input and generates a phoneme sequence as the output. In practice, the grapheme sequence is converted into a grapheme embedding sequence via an embedding table. Considering the order of the input and output sequence, the positional encoding [48] is combined with the input of both the G2P encoder and the G2P decoder to provide the positional information within a sequence. The G2P encoder consists of a stack of 3 identical Transformer encoders. Each Transformer encoder has a multi-head self-attention layer and a fully connected feed-forward network. The G2P encoder maps the grapheme embedding with the positional encoding to a sequence of latent textual representations that is the same length as the input grapheme sequence.

The G2P decoder consists of a stack of 3 identical Transformer decoders. Each Transformer decoder has a masked multi-head self-attention layer and a fully connected feed-forward network. In the middle of these two sub-layers, there is a multi-head attention layer to learn the alignment between the grapheme sequence and the phoneme sequence. The attention mechanism for alignment prediction uses the output of the masked multi-head self-attention module and attends to the sequential latent textual representations from the G2P encoder. For each accent, we have an accented phonetic lexicon. An English grapheme sequence can have multiple phonetic transcriptions, one for each accent, as the example shown in Section I. To control the accent variation during training of the multi-accent G2P model, we use an accent ID as an additional input to the decoder. The accent ID guides the G2P model to generate the accented phoneme sequence from the same accented lexicon. The phoneme sequence and the accent ID are converted to the phoneme embedding and the accent embedding, via two different embedding tables in the pre-net, respectively. The accent embedding is a vector that applies to the entire phoneme sequence. Specifically, the accent embedding is duplicated at the phoneme-level and concatenated with the phoneme embedding. The combination of phoneme embedding and accent embedding is then passed to a linear projection layer in the pre-net. The G2P decoder takes the encoded latent textual representation and the output of the pre-net together with the positional encoding to generate phonetic output embeddings in an autoregressive manner, using the output phoneme of the previous step as the input at the current step.

The embedding vectors from the decoder output are converted to the discrete phoneme sequence by a linear projection layer with a softmax function. The softmax function produces the probability that the predicted phoneme belongs to each phoneme class. At run-time, the phoneme class that has the highest probability is predicted. The G2P decoder works on the principle of teacher-forcing. During training, it takes the ground truth phoneme sequence as input, while at run-time, it takes the predicted phoneme sequence as the contextual input.

The loss function for training the G2P model is defined as the cross-entropy (CE) between the G2P output and the ground

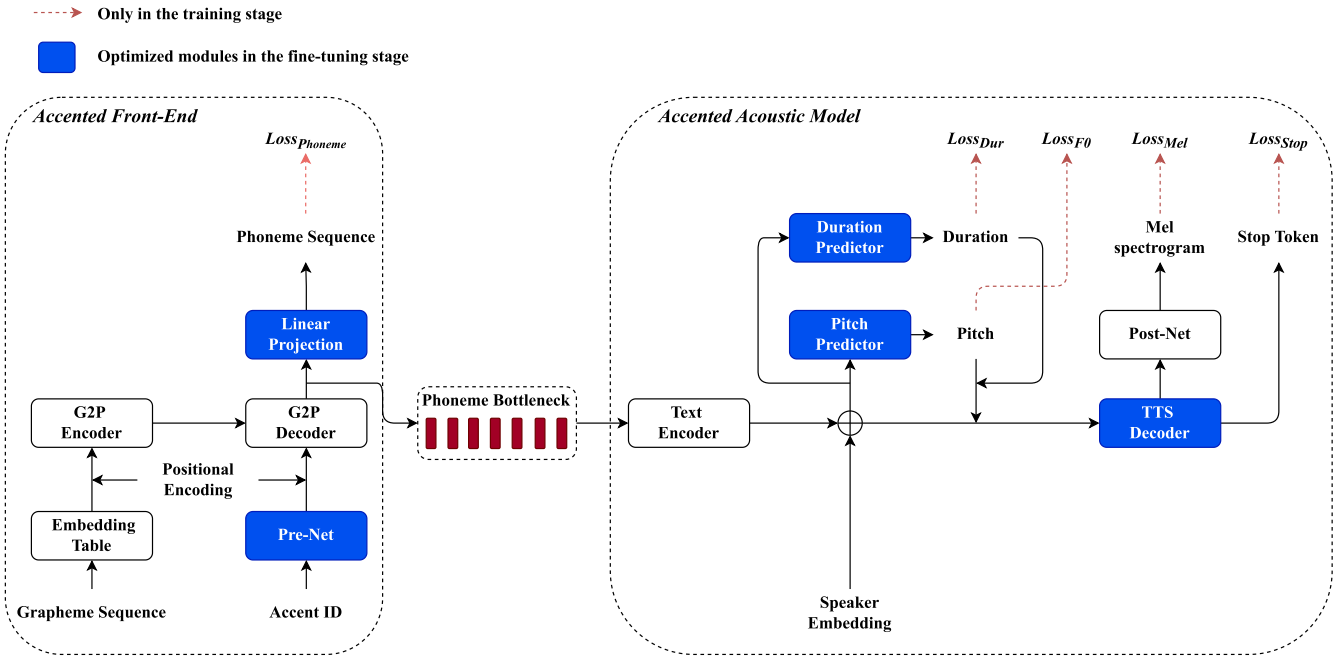


Fig. 2. Block diagram of the proposed accented TTS framework. The left panel is the accented front-end for G2P conversion, while the right panel is the accented acoustic model for phoneme-to-Mel-spectrogram prediction. The phoneme bottleneck extracted from the G2P decoder output serves as the input of the acoustic model encoder. The front-end and acoustic model are pre-trained and then fine-tuned individually. All modules shown in this figure are involved during the pre-training stage. During the fine-tuning stage, only the parameters in the blue color modules are updated, while the other modules in the white color are fixed. The speaker encoder and neural vocoder are omitted for simplicity.

truth phoneme sequence:

$$Loss_{CE} = -\sum_{i=1}^{N} log\left(p\left(x_i|\hat{x}_i\right)\right) \tag{1}$$

where $N$ is the number of total classes, and $p(x_i|\widehat{x}_i)$ is the output of the softmax function, i.e., the probability that the predicted phoneme $\widehat{x}_i$ belongs to the corresponding phoneme label $x_i$.

*2) Accented Acoustic Model:* The right panel of Fig. 2 is a variant of TTS encoder-decoder acoustic model. In a standard TTS system, an embedding table maps a phoneme to its embedding vector in a continuous space. However, some phonemes in one accent may not exist in another accent. To generalize an accented front-end trained for one accent to another, we propose to make use of the embedding vector from the G2P decoder output, i.e., the phoneme bottleneck, instead of the discrete phoneme ID as the input to the TTS acoustic model. In this way, we hope to handle the pronunciation of unseen phonemes, even when the speech training data for the target accent is not available. This will be discussed in Section V-A.2).

To facilitate the fine-tuning with limited accented speech data for the target accent, we propose to pre-train a multi-speaker TTS acoustic model. The proposed acoustic model also follows the strategy of pre-training and fine-tuning. It adopts the Tacotron 2 architecture, a data-driven model that represents the state-of-the-art performance in TTS. However, Tacotron 2 does not explicitly model the prosodic attributes of accents, e.g., pitch pattern and phoneme duration. Therefore, we incorporate pitch and duration predictors into the TTS acoustic model to enable the model to predict accurate pitch and phoneme duration that are close to ground truth.

As shown in Fig. 2, the TTS acoustic model takes phoneme bottleneck features as input. The text encoder is composed of 3 convolutional neural networks (CNNs) and a bidirectional long short-term memory (LSTM). The text encoder converts phoneme bottleneck features into the sequential hidden phoneme representation. We condition the multi-speaker TTS pre-training on a $d$-vector-based speaker embedding, which represents the speaker identity. Since the pitch pattern and phoneme duration are also affected by the individual speaker, the speaker embedding is added to the output of the text encoder before the pitch and duration predictors. Phoneme modeling handled by the text encoder is an essential part of accented speech representation. Different from FastSpeech 2 [8], in this work, we adopt both duration predictor and attention mechanism to provide more accurate phoneme boundary with limited speech data, and the pitch and duration predictors are incorporated with the text encoder. The architectures of the pitch and duration predictors are the same and consist of two 1D CNNs and a linear projection layer. They use the combination of the output of the text encoder and the speaker embedding to predict the $F0$ and duration at the phoneme-level, respectively. The $F0$ and phoneme duration are added with the TTS acoustic model via an embedding table and a linear projection layer, respectively.

The decoder network takes the combination of text encoder output, speaker embedding, $F0$, and phoneme duration to predict Mel-spectrogram and the stop token label. The location-sensitive attention mechanism in the decoder learns the duration alignment between the phoneme and the Mel-spectrogram. There are two fully connected layers, two LSTM layers, and two linear projection layers in the decoder to predict the Mel-spectrogram and the stop token label in an autoregressive

manner [5]. A post-net with the residual connection is used for the Mel-spectrogram reconstruction.

In the training stage, the ground truth of the Mel-spectrogram, $F0$, and phoneme duration are used in a teacher-forcing manner, while the predicted ones are used at run-time. The stop token loss function is defined as the binary cross-entropy, and the mean squared error (MSE) is adopted as the loss function for the prediction of Mel-spectrogram, $F0$, and phoneme duration:

$$Loss_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x_i})^2 \tag{2}$$

where $N$ is the length of the sequence, $x_i$ and $\widehat{x_i}$ are the values from the ground truth sequence and predicted one at the $i$th element.

The total loss function of the proposed TTS acoustic model is defined as:

$$\begin{aligned} Loss_{\text{Total}} &= \alpha Loss_{Mel} + \beta Loss_{Stop} \\ &\quad + \gamma Loss_{F0} + \delta Loss_{Dur} \end{aligned} \tag{3}$$

where $Loss_{Mel}$ is the summarization of the Mel-spectrogram reconstruction losses before and after the post-net.

### B. Accented TTS Training With Limited Data

*1) Only a Small Accented Phonetic Lexicon:* We create an accented front-end for the target accent to address the phonetic variation problem with only a small target accented phonetic lexicon. Specifically, we fine-tune the accent-related layers of the pre-trained multi-accent G2P model. The accent-related layers are the layers related to phoneme generation. The fine-tuned modules are drawn in blue, as shown in the Accented Front-End model in Fig. 2.

In an early study [43], a low-resource G2P model with high performance is built by fine-tuning only the phoneme embedding table and the linear projection layer. Similarly, we use the pre-trained model as initialization and freeze the grapheme embedding table, G2P encoder, and G2P decoder modules, which are shared for all accents during fine-tuning. In [43], the phoneme embedding table and the linear projection layer are re-initialized in the fine-tuning stage, because the phoneme symbols are different in their pre-training and fine-tuning stages. In contrast, we use the unified phoneme symbols for all accents used in pre-training and fine-tuning stages. Therefore, instead of the re-initialization, we fine-tune the following accent-related layers based on their pre-trained parameters: the accent embedding table and the phoneme embedding table in the pre-net, and the linear projection layer after the G2P decoder. These modules are fine-tuned so that the G2P model generates the phoneme sequence of the target accented phonetic lexicon.

*2) Both a Small Accented Phonetic Lexicon and Limited Accented Speech Samples:* Besides fine-tuning the pre-trained G2P model, we further address the prosodic variation in the generated accented speech by utilizing the limited accented speech samples for the target accent. Specifically, we use the pre-trained TTS acoustic model as initialization and fine-tune the accent-related layers to modify the prosodic components. As shown in Fig. 2, the fine-tuned modules of the Accented Acoustic Model are drawn in blue color.

In a previous study [49], it has been reported that fine-tuning the text encoder with limited speech data could degrade the TTS performance due to the uneven distribution of linguistic information. Therefore, we freeze the text encoder in the fine-tuning stage. The attention layer and decoder aim to learn phoneme alignment and Mel-spectrogram prediction. To reduce the gap between the pre-trained acoustic model and the target accent, we fine-tune the attention layer and decoder. Since the pitch and duration predictors explicitly control the $F0$ and phoneme duration of the generated speech, we fine-tune both predictors to provide the information regarding the prosodic variation for the target accent. In this way, the acoustic model is able to generate accented speech with more accurate pitch and phoneme duration. The generalization ability of the acoustic feature reconstruction of the post-net can be improved using a large number of Mel-spectrogram in the pre-training stage. Thus, the pre-trained post-net is fixed and shared for all accents in the fine-tuning stage. All the loss functions in (3) are optimized in the fine-tuning stage.

## IV. Experimental Setup

We conduct experiments to validate the effectiveness of the proposed accented front-end model and accented acoustic model on accent rendering. Both models are first pre-trained on a large amount of data and then fine-tuned on limited accented data.

### A. Accented Front-End

All phonetic lexicons involved are from the Unisyn Lexicon [50], which covers several accents of English with a unified phoneme symbol inventory. If a word in a lexicon has multiple phonetic transcriptions, we take the first one. We first pre-train a multi-accent G2P model using accented phonetic lexicons of General American English, Irish English, and British English Received Pronunciation with 56, 59, and 50 phonemes, respectively. We then fine-tune the pre-trained G2P model separately using accented phonetic lexicons of Scottish and General Australian English with 61 and 51 phonemes, respectively. In the fine-tuning stage, we vary the lexicon size of 1 k, 5 k, 10 k, and 40 k most frequent words, in addition to the full lexicon size of about 120 k words to observe the effect. When training the G2P model, the input is one utterance rather than a single word as in [43]. The text transcripts are selected from the LibriTTS [51] corpus. The utterances that are too short or too long and the utterances with any out-of-vocabulary word according to the phonetic lexicon are removed from the dataset.

Two front-end models are implemented for comparison:

- *SA-G2P:* This is a pre-trained single-accent G2P model trained only with General American English.
- *MA-G2P:* This is a pre-trained multi-accent G2P model trained with multiple accented phonetic lexicons including General American English, Irish English, and British English Received Pronunciation. The accent ID is used as an additional input in the multi-accent G2P model.

TABLE I
SUMMARY OF THE DATABASE USED FOR BUILDING THE ACCENTED ACOUSTIC MODEL

| | Pre-Train | Fine-Tune | |
|---|---|---|---|
| Database | VCTK | CMU_ARCTIC | Google_TTS_API |
| # of Accent | 3 | 1 | 1 |
| # of Speaker | 45 | 1 | 1 |
| # of Utterance | 16,600 | 50 - 300 | 50 - 300 |
| Total duration | 9.66 hours | 2.63 - 14.96 mins | 3.10 - 18.29 mins |

The model architectures and network configurations of the Transformer-based G2P model are shown in Table II. Each G2P model is pre-trained for 100 epochs and fine-tuned for another 50 epochs for each experimental group. The model with the lowest validation loss among pre-training epochs is used as the initialization during the fine-tuning stage, and the model with the lowest validation loss among fine-tuning epochs is used for testing. The learning rate is $5e$-4 and the batch size is 128 in both pre-training and fine-tuning stages.

### B. Accented Acoustic Model

We first pre-train a multi-speaker TTS acoustic model using the subset of the CSTR_VCTK [52] speech corpus. To capture the shared knowledge across accents, we pre-train the model with speech data from multiple accents. We select the dataset with the same accents as we pre-train the G2P model, i.e., the accents of General American English, Irish English, and British English Received Pronunciation. Then, we fine-tune the pre-trained acoustic model with a Scottish speaker's data from CMU_ARCTIC [53] corpus, and an Australian speaker's data generated by Google TTS API[1], respectively. We randomly select 50, 100, 200, and 300 utterances as the training data during the fine-tuning stage. The details of the involved database are summarized in Table I. All speech signals are down-sampled to 16 $k$ Hz and the silence at the beginning and end of each utterance is trimmed. We use the logarithmic scale 80-dim Mel-spectrogram as the acoustic feature that is extracted with 12.5 ms frame-shift and 50 ms frame length.

For the training of the phoneme-level pitch and duration predictors, we obtain the phoneme boundary by applying force-alignment using an automatic speech recognition (ASR) system. As a generic ASR model does not work well across accents, we train the accent-dependent ASR for force-alignment.[2] The phoneme duration is represented as the number of frames belonging to the same phoneme and transformed into a logarithmic scale. The $F0$ is extracted using pyworld[3] with 12.5 ms frame shift. The linear interpolation method is adopted on the unvoiced frames of $F0$. The frame-level $F0$ is down-sampled to the phoneme-level according to the phoneme duration. Specifically, we take the average of $F0$ values on the frames belonging to the same phoneme. The phoneme-level $F0$ is then normalized to have zero mean and unit variance over the speech data for pre-training. The 256-dim speaker embedding is extracted from a pre-trained speaker encoder. The speaker encoder is trained with AISHELL-2 dataset [54], following [55].

The following six accented TTS systems are implemented for comparison:

- *Char-AM:* This is a multi-speaker TTS acoustic model that takes a character sequence as input.
- *US_G2P-AM:* This is an accented TTS framework that consists of an American G2P model and a multi-speaker TTS acoustic model. The American G2P model is part of the pre-trained multi-accent G2P model conditioned on the General American accent ID.
- *SCOT_G2P-AM:* This is an accented TTS framework that consists of a Scottish G2P model and a multi-speaker TTS acoustic model. The Scottish G2P model is the pre-trained multi-accent G2P model fine-tuned with a Scottish phonetic lexicon of 5 $k$ words.
- *SCOT_G2P(PS)-AM:* This is similar to SCOT_G2P-AM except that the input for the TTS acoustic model is discrete phoneme ID sequence from G2P output.
- *AU_G2P-AM:* This is similar to SCOT_G2P-AM except that the G2P model is fine-tuned with a General Australian phonetic lexicon of 5 $k$ words.
- *AU_G2P(PS)-AM:* This is similar to SCOT_G2P(PS)-AM except that the G2P model is fine-tuned with a General Australian phonetic lexicon of 5 $k$ words.
- *SCOT_G2P-F0_Dur_AM:* This is an accented TTS framework that consists of a Scottish G2P model and a multi-speaker TTS acoustic model with integrated pitch and duration predictors. The Scottish G2P model is the pre-trained multi-accent G2P model fine-tuned with a Scottish phonetic lexicon of 5 $k$ words.
- *AU_G2P-F0_Dur_AM:* This is similar to SCOT_G2P-F0_Dur_AM except that the G2P model is fine-tuned with a General Australian phonetic lexicon of 5 $k$ words.

All G2P models and acoustic models involved are pre-trained. In the following experiments, the TTS system followed with '-L' denotes fine-tuning the G2P model with only a small accented phonetic lexicon, with '-S' denotes fine-tuning the acoustic model with only limited accented speech data, and with '-LS' denotes using both of them. If neither of these options is used, it is a pre-trained system. Note that the TTS system with (PS) takes the phoneme sequence as acoustic model input, otherwise the phoneme bottleneck is taken as input to the text encoder.

The model architectures and network configurations of the Tacotron 2-based acoustic model are summarized in Table II. Each of the TTS acoustic models is pre-trained for 800 epochs and fine-tuned for another 100 epochs for each experimental group using Adam optimizer [56]. All weights $\alpha$, $\beta$, $\gamma$, and $\delta$ in (3) are set to 1. The learning rate is $1e$-3 and $1e$-4, and the batch size is 32 and 8 for the pre-training stage and fine-tuning stage, respectively.

### C. Waveform Generation

We select Parallel WaveGAN [40] as the neural vocoder to reconstruct the time-domain speech waveform from the predicted

[1][Online]. Available: https://cloud.google.com/text-to-speech
[2][Online]. Available: https://montreal-forced-aligner.readthedocs.io/en/latest
[3][Online]. Available: https://pypi.org/project/pyworld

TABLE II
SUMMARY OF MODEL ARCHITECTURES AND NETWORK CONFIGURATIONS FOR THE PROPOSED ACCENTED TTS FRAMEWORK THAT CONSISTS OF AN ACCENTED G2P MODEL AND AN ACCENTED TTS ACOUSTIC MODEL

| | |
|---|---|
| G2P_Embedding | Grapheme embedding table: 256-D<br>Positional encoder: 256-D → Dropout(0.1) |
| G2P_Encoder | Encoder stack number: 3<br>Multi-head attention: 8 heads of 256-D<br>→ Dropout(0.1) → LN<br>Feedforward: FC-512-ReLU → Dropout(0.1)<br>→ FC-256 → Dropout(0.1) → LN |
| G2P_Attention | Multi-head attention: 8 heads of 256-D<br>→ Dropout(0.1) → LN |
| G2P_Pre-Net | Phoneme embedding table: 256-D<br>Accent embedding table: 32-D<br>Linear projection: FC-256<br>Positional encoding: 256-D → Dropout(0.1) |
| G2P_Decoder | Decoder stack number: 3<br>Masked multi-head attention: 8 heads of 256-D<br>→ Dropout(0.1) → LN<br>Feedforward: FC-512-ReLU → Dropout(0.1)<br>→ FC-256 → Dropout(0.1) → LN |
| G2P_Projection | FC-82 |
| TTS_Encoder | Phone linear projection: FC-512-softsign<br>3 layers of conv1d-5-512-BN1d<br>→ Bi-LSTM-512<br>Speaker embedding projection: FC-512-softsign |
| TTS_Attention | Location layer: conv1d-31-32<br>→ Linear projection: FC-128 |
| TTS_Decoder | Pre-net: FC-256-ReLU → Dropout(0.5)<br>→ FC-256-ReLU → Dropout(0.5)<br>RNN: LSTM-1024 → Dropout(0.1)<br>→ LSTM-1024 → Dropout(0.1)<br>Mel linear projection: FC-80<br>Stop token linear projection: FC-1 |
| TTS_Post-Net | 4 layers of conv1d-5-512-BN1d<br>→ conv1d-5-80-BN1d |
| Pitch Predictor | Conv1d-3-512-ReLU-LN → Dropout(0.5)<br>→ Conv1d-3-256-ReLU-LN → Dropout(0.5)<br>→ Linear projection: FC-1<br>Pitch embedding table: 512-D |
| Duration Predictor | Conv1d-3-512-ReLU-LN → Dropout(0.5)<br>→ Conv1d-3-256-ReLU-LN → Dropout(0.5)<br>→ Linear projection: FC-1<br>Duration projection: FC-512-ReLU → Dropout(0.5) |
| Speaker Encoder | 3 layers of LSTM-768<br>→ Linear projection: FC-256 |

FC-X denotes a fully connected layer with X units. Conv1D-K-C means 1-D convolution with width K and C output channels. LN denotes layer normalization, and BN indicates batch normalization.

Mel-spectrogram for all experiments owing to its capability of generating high-quality speech waveform with a rapid speed. It is pre-trained with the CSTR_VCTK [52] corpus, with the same frame-shift and frame length as the TTS acoustic model.

## V. EXPERIMENTAL ANALYSIS

We conduct both objective and subjective evaluations. In this section, we report the experiments under two scenarios, a) only a small accented phonetic lexicon is available. b) both a small accented phonetic lexicon and limited accented speech samples are available.

### A. Only a Small Accented Phonetic Lexicon

In this scenario, we only fine-tune the pre-trained multi-accent G2P model, while the pre-trained multi-speaker TTS acoustic model is frozen. We evaluate the phonetic variation in the accent rendering in terms of the accuracy of the output phoneme of front-end model and the output speech of the complete system.

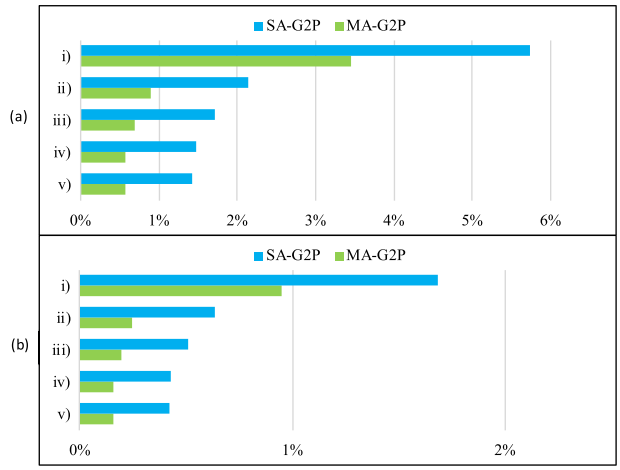


Fig. 3. Performance of both MA-G2P and SA-G2P fine-tuned with a Scottish phonetic lexicon of difference sizes in terms of WER and PER. (i)–(v) indicate the lexicon of 1 k, 5 k, 10 k, 40 k words and full lexicon size, respectively. (a) WER (b) PER.

*1) Objective Evaluation:* We report the performance of the G2P model in terms of phoneme error rate (PER) and word error rate (WER). PER and WER indicate the Levenshtein distance between the predicted sequence and reference at the phoneme and word level, respectively. Lower PER and WER account for more accurate predicted phoneme sequences. To evaluate the phonetic variation of the output speech of TTS system, we calculate the Kullback-Leibler divergence [57] (KLD) of phonetic posteriorgram between the generated speech and reference speech to measure the phonetic distribution similarity. The lower value suggests a better performance.

*i) Accuracy of Accented Front-End:* We compare two G2P models: accented G2P by fine-tuning MA-G2P and that by fine-tuning SA-G2P [43] on lexicons of different sizes. The results on Scottish and Australian accents are illustrated in Figs. 3 and 4, respectively, and show that the fine-tuned MA-G2P outperforms the fine-tuned SA-G2P across all test cases. The results are also consistent between Scottish and Australian accents.

It suggests that MA-G2P benefits from shared phonetic knowledge across accents. It is also observed that the Scottish G2P works better than the Australian counterpart, which can be explained by the fact that Scottish English accent is closer to Irish English and British English Received Pronunciation than the Australian English accent.

As fine-tuning MA-G2P with a lexicon of 5 $k$ words achieves a reasonable performance (PER 0.25% and WER 0.90% on Scottish accent, PER 0.46% and WER 1.46% on Australian accent) that is comparable with full lexicon size (PER 0.16% and WER 0.58% on Scottish accent, PER 0.33% and WER 1.08% on Australian accent), we select the G2P fine-tuned with an accented phonetic lexicon of 5 $k$ words as our accented front-end in the limited data scenario for all subsequent experiments.

*ii) Phonetic Variation:* We perform statistical analysis on accented phonetic lexicons to understand the difference between

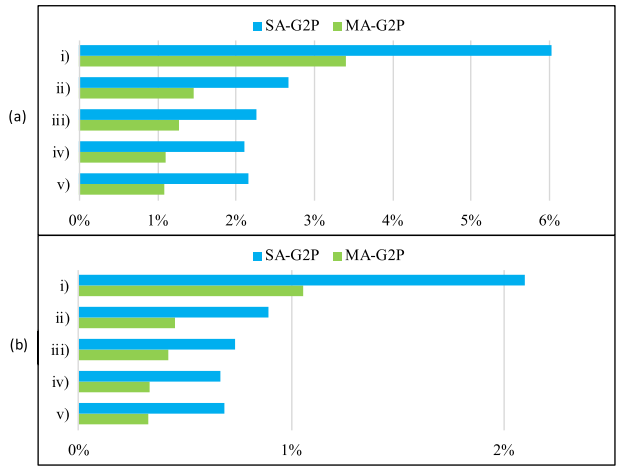


Fig. 4. Performance of both MA-G2P and SA-G2P fine-tuned with a General Australian phonetic lexicon of difference sizes in terms of WER and PER. (i)–(v) indicate the lexicon of 1 k, 5 k, 10 k, 40 k words and full lexicon size, respectively. (a) WER (b) PER.

TABLE III
COMPARISON OF PHONETIC TRANSCRIPTIONS BETWEEN SCOTTISH / GENERAL AUSTRALIAN ACCENT AND GENERAL AMERICAN ACCENT

| | Scottish | Australian |
|---|---|---|
| Shared Word (%) | 26.62 | 22.22 |
| Accented Word (%) | 73.38 | 77.78 |
| Vowel Variation (%) | 24.07 | 41.16 |
| Consonant Variation (%) | 22.04 | 19.66 |

Scottish / General Australian accent and General American accent. We count the percentage of the shared words whose phonetic transcriptions are the same between two lexicons and the accented words whose phonetic transcriptions are different between two lexicons. We further analyze the accented words in terms of vowel and consonant variations. Specifically, we calculate the rate of vowel and consonant differences by the Levenshtein distance on the accented words between two lexicons, respectively. As shown in Table III, the vowel variation is more prominent than the consonant one. The same findings have been reported in prior studies [22], [58] on accented speech corpora.

We therefore evaluate accent rendering by focusing on the vowels on the accented words. We select 191 accented words with a total of 304 vowel variations from the test set on Scottish accent. We extract the phonetic posteriorgram from a pre-trained speaker-independent ASR acoustic model. The frame-level posteriorgram is then down-sampled to the phoneme-level according to the phoneme boundary obtained by force-alignment. We calculate the KLD of the phoneme-level posteriorgram to compare the vowel distribution similarity, as shown in Table IV. We can observe that the generated speech from SCOT_G2P-AM-L has a closer vowel distribution to reference speech than that from US_G2P-AM on Scottish accent. We also observe that both G2P-AM outperform Char-AM.

From the above observations, we could claim the following two statements: (a) TTS system with phoneme input is more able to model phonetic variation accurately than that with character input. (b) The pre-trained G2P fine-tuned with a small accented phonetic lexicon effectively changes the phonetic variation of output speech of TTS system.

TABLE IV
KL DIVERGENCE (KLD) OF VOWEL DISTRIBUTION ON THE ACCENTED WORDS FOR THREE COMPARATIVE TTS SYSTEMS ON SCOTTISH ACCENT

| System | KLD |
|---|---|
| Char-AM | 6.60 |
| US_G2P-AM | 5.36 |
| SCOT_G2P-AM-L | **5.15** |

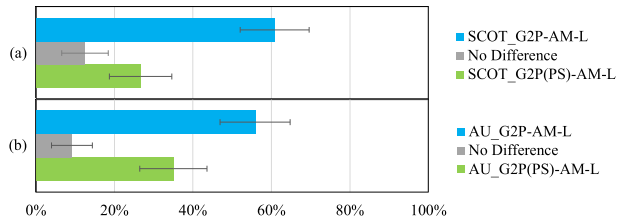


Fig. 5. AB test results for naturalness of accented TTS systems with different input representations from G2P output, discrete phoneme ID versus the phoneme bottleneck, on both Scottish and Australian accents with 95% confidence intervals. (a) Scottish accent. (b) Australian accent.

*2) Subjective Evaluation:* The subjective evaluations are conducted through listening tests by human subjects. 20 listeners who are proficient in English participate in each experimental set of the listening tests.[4] We conduct AB preference test to evaluate the speech quality in terms of naturalness, and XAB preference test to assess the accent pronunciation similarity.

*i) Speech Quality:* In the AB test, participants are asked to listen to the two compared samples and choose the better one based on naturalness. Each listener listens to 12 samples in a single experiment, i.e., a total of 24 ($12 \times 2$ (# of accents) $=$ 24) samples. The results are shown in Fig. 5. SCOT_G2P-AM-L and AU_G2P-AM-L with phoneme bottleneck input are compared with their systems with phoneme ID input. We clearly observe that accented TTS system with continuous phoneme representation as input synthesizes accented speech with better speech quality than the system with discrete phoneme token as input. The target accented phonemes from the fine-tuned G2P contain the unseen phonemes compared to the accents used in the pre-training phase. Since only a small lexicon is available, for the acoustic model with discrete phoneme ID as input, the pronunciation of the unseen phonemes degrades the speech quality of the generated accented speech. Therefore, in this work we use the phoneme bottleneck as input for our acoustic model.

*ii) Accent Pronunciation Similarity:* In the XAB test, we label the accented words in each utterance in red color. The listeners are required to only pay attention to the labeled accented words and ignore other words. Each listener listens to the reference speech first and then chooses a more similar pronunciation to the reference speech from two different samples. Each listener listens to 18 samples in a single experiment, thus in a total of 36 ($18 \times 2$ (# of accents) $=$ 36) samples. As shown in

[4]All speech samples are available at: https://xuehao-marker.github.io/taslp_G2P-TTS/

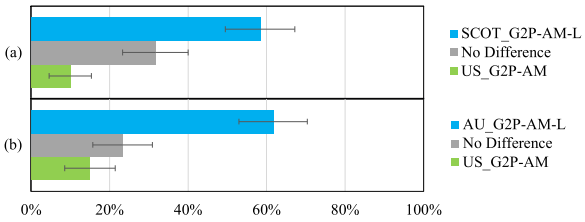


Fig. 6. XAB test results for accent pronunciation similarity of accented TTS systems with G2P pre-trained and fine-tuned with a small accented phonetic lexicon on both Scottish and Australian accents with 95% confidence intervals. (a) Scottish accent. (b) Australian accent.

Fig. 6, US_G2P-AM is compared with SCOT_G2P-AM-L to show the effect of the accented G2P on Scottish accent. Also, US_G2P-AM is compared with AU_G2P-AM-L on Australian accent. It can be seen that the TTS system with fine-tuned G2P significantly achieves better performance than that with pre-trained US_G2P in terms of accent pronunciation similarity, which is consistent on both Scottish and Australian accents. The results strongly support the idea of accented G2P modeling on addressing phonetic variation problem, that only depends on a small accented phonetic lexicon of 5 $k$ words without the need for accented speech samples during training.

### *B. Both a Small Accented Phonetic Lexicon and Limited Accented Speech Samples*

After studying the effect of fine-tuning a pre-trained G2P model with a small accented phonetic lexicon, we further fine-tune the TTS acoustic model with a limited amount of accented speech data, to study the impact of the prosodic variation on accent rendering. We evaluate the performance on Scottish and Australian accents in terms of pitch pattern, phoneme duration, speech quality, and accent similarity. All experimental systems are tested on the same 100 unseen utterances.

*1) Objective Evaluation:* We calculate the $F0$ root mean squared error (RMSE), logarithmic scale $F0$ correlation coefficient [59], and unvoiced/voiced (U/V) error rate to evaluate the pitch amplitude and trajectory trend similarity. We calculate the frame disturbance [60] and phoneme duration RMSE to evaluate the accuracy of the predicted duration. Dynamic time warping (DTW) [61] is used to align the generated Mel-spectrogram and the reference. To further demonstrate the effectiveness of pitch and duration predictors on the fine-tuning with limited accented speech data, we fine-tune the TTS acoustic model with 50, 100, 200, and 300 utterances, respectively. The average objective test results are summarized in Table V.

*i) Pitch:* The $F0$ RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \widehat{x}_i)^2} \tag{4}$$

where $N$ is the total number of frames, $x_i$ and $\widehat{x}_i$ are the $F0$ values from the reference speech and generated speech at the $i$th frame. The lower $F0$ RMSE indicates the lower pitch amplitude error. Instead of log $F0$ RMSE [62], we calculate the original $F0$ RMSE over the both voiced frames between the aligned generated speech and reference speech.

The log-scale $F0$ correlation coefficient is defined as:

$$C = \frac{\sum_{i=1}^{N}(y_i - \overline{y})\left(\widehat{y}_i - \overline{\widehat{y}}\right)}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{N}\left(\widehat{y}_i - \overline{\widehat{y}}\right)^2}} \tag{5}$$

where $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\overline{\widehat{y}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{y}_i$, $y_i$ and $\widehat{y}_i$ are the log-scale $F0$ values from the reference speech and generated speech at the $i$th frame and $N$ is the number of the calculation frames. We calculate the correlation coefficient only on both voiced frames of the $F0$ from aligned reference speech and generated speech. The correlation coefficient value is between the interval of $[-1, 1]$ and the closer to 1 indicates the higher pitch trajectory trend similarity.

We compare the U/V labels between the generated speech and reference speech in terms of U/V error rate, which is the ratio of the frame discrepancy count to the total number of frames. A lower U/V error rate indicates a better pitch reconstruction.

In Table V, we make three observations, (a) All TTS systems fine-tuned with accented speech data outperform those without. This confirms the effectiveness of prosody modeling in the TTS acoustic model. It also suggests that the unseen $d$-vector-based speaker embedding is able to represent the speaker characteristic, while it lacks accent-related prosodic information. (b) From the four fine-tuning cases with different amounts of accented speech data, we note that the pitch performance of all G2P-AM-LS is better than that of Char-AM-S. This confirms the contribution of the G2P model to prosodic rendering. (c) We are glad to see that our proposed G2P-F0_Dur_AM-LS consistently achieves the lowest $F0$ RMSE, the highest correlation, and the lowest U/V error rate among the three compared systems in each fine-tuning case. This observation strongly demonstrates that the pitch predictor contributes to the more accurate pitch trajectory in the generated speech with only limited speech data for the target accent. The observations are consistent on both Scottish and Australian test cases.

We further visualize the pitch ($F0$) contours generated by the three comparative systems and the reference speech. The $F0$ contours on one Scottish utterance are shown in Fig. 7. We observe that the $F0$ contour of SCOT_G2P-F0_Dur_AM-LS is consistently the most similar one to the reference $F0$ across all systems, and collaborates the objective evaluation results in Table V.

*ii) Duration:* The frame disturbance is defined as:

$$Disturbance = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_t - \widehat{p}_t)^2} \tag{6}$$

where $p_t$ and $\widehat{p}_t$ are the aligned path between the reference and generated speech at the $i$th frame, and $N$ is the number of the aligned frames. The lower disturbance value indicates the smaller duration distortion between the reference speech and generated speech.

The definition of phoneme duration RMSE is similar to (4), except that $x_i$ and $\widehat{x}_i$ here are the phoneme duration of the reference and generated speech for the $i$th phoneme and $N$ is the

TABLE V
OBJECTIVE EVALUATION RESULTS IN TERMS OF PITCH AND DURATION OF THE COMPARATIVE TTS SYSTEMS FINE-TUNED WITH DIFFERENT SIZES OF ACCENTED SPEECH DATA ON SCOTTISH AND AUSTRALIAN ACCENTS

| Accent | # of Utt / Duration | System | Pitch | | | Duration | |
|---|---|---|---|---|---|---|---|
| | | | RMSE(Hz) | Correlation | U/V(%) | Disturbance(frame) | RMSE(ms) |
| Scottish | - | Char-AM | 30.21 | 0.67 | 16.35 | 21.53 | 45.48 |
| | | SCOT_G2P-AM-L | 35.33 | 0.67 | 15.27 | 19.18 | 51.49 |
| | | SCOT_G2P-$F0$_Dur_AM-L | 43.57 | 0.68 | 14.98 | 12.34 | 46.94 |
| | 50 / 2.63 mins | Char-AM-S | 20.09 | 0.75 | 14.08 | 13.10 | 47.40 |
| | | SCOT_G2P-AM-LS | 19.78 | 0.77 | 12.78 | 10.51 | 44.55 |
| | | SCOT_G2P-$F0$_Dur_AM-LS | **17.47** | **0.81** | **11.61** | **10.24** | **42.66** |
| | 100 / 5.03 mins | Char-AM-S | 17.74 | 0.78 | 13.10 | 13.04 | 43.90 |
| | | SCOT_G2P-AM-LS | 17.62 | 0.80 | 11.88 | 10.13 | 42.21 |
| | | SCOT_G2P-$F0$_Dur_AM-LS | **16.64** | **0.83** | **11.63** | **9.85** | **40.79** |
| | 200 / 9.93 mins | Char-AM-S | 16.85 | 0.74 | 12.93 | 13.25 | 45.22 |
| | | SCOT_G2P-AM-LS | 16.40 | 0.80 | 12.00 | 10.15 | **40.46** |
| | | SCOT_G2P-$F0$_Dur_AM-LS | **12.92** | **0.82** | **11.28** | **9.50** | 41.31 |
| | 300 / 14.96 mins | Char-AM-S | 15.93 | 0.78 | 12.66 | 12.89 | 44.02 |
| | | SCOT_G2P-AM-LS | 15.51 | 0.81 | 11.83 | 10.04 | 40.88 |
| | | SCOT_G2P-$F0$_Dur_AM-LS | **12.26** | **0.82** | **10.45** | **9.47** | **40.60** |
| Australian | - | Char-AM | 40.88 | 0.37 | 19.04 | 12.85 | 40.11 |
| | | AU_G2P-AM-L | 42.24 | 0.47 | 17.03 | 13.44 | 36.36 |
| | | AU_G2P-$F0$_Dur_AM-L | 40.48 | 0.54 | 17.73 | 9.97 | 41.01 |
| | 50 / 3.10 mins | Char-AM-S | 36.44 | 0.67 | 16.71 | 12.26 | 31.96 |
| | | AU_G2P-AM-LS | 31.91 | 0.73 | 15.25 | 11.30 | 30.05 |
| | | AU_G2P-$F0$_Dur_AM-LS | **29.02** | **0.78** | **15.12** | **7.13** | **28.98** |
| | 100 / 5.94 mins | Char-AM-S | 35.07 | 0.70 | 16.41 | 11.59 | 30.38 |
| | | AU_G2P-AM-LS | 30.91 | 0.76 | 15.33 | 7.32 | 29.75 |
| | | AU_G2P-$F0$_Dur_AM-LS | **27.43** | **0.81** | **15.09** | **6.28** | **28.74** |
| | 200 / 12.21 mins | Char-AM-S | 33.49 | 0.73 | 16.20 | 12.03 | 28.02 |
| | | AU_G2P-AM-LS | 27.97 | 0.80 | 14.94 | 6.74 | 27.31 |
| | | AU_G2P-$F0$_Dur_AM-LS | **26.42** | **0.83** | **14.67** | **5.75** | **26.93** |
| | 300 / 18.29 mins | Char-AM-S | 31.76 | 0.75 | 16.10 | 11.15 | 28.00 |
| | | AU_G2P-AM-LS | 27.17 | 0.81 | 14.73 | 6.83 | 27.02 |
| | | AU_G2P-$F0$_Dur_AM-LS | **25.58** | **0.83** | **14.57** | **5.47** | **26.39** |
| Average | - | Char-AM-S | 25.92 | 0.74 | 14.77 | 12.42 | 37.36 |
| | | G2P-AM-LS | 23.41 | 0.78 | 13.59 | 9.13 | 35.28 |
| | | G2P-$F0$_Dur_AM-LS | **20.97** | **0.82** | **13.05** | **7.96** | **34.55** |

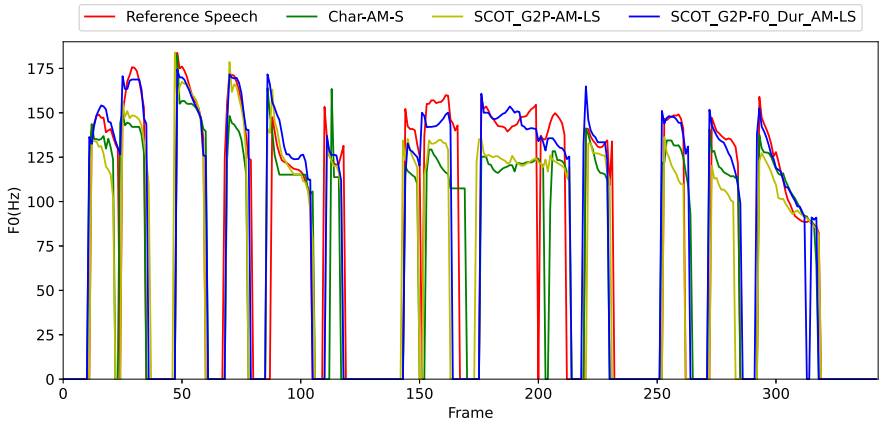


Fig. 7. $F0$ contours on one Scottish utterance of the text transcription "I just do appreciate it without being able to express my feelings". They are extracted from the reference speech and generated speech from three comparison systems fine-tuned with 300 accented utterances.

number of phonemes. A lower duration RMSE indicates better phoneme duration reconstruction.

In the last two columns of Table V, we report the duration evaluation results on Scottish and Australian accents. We observe the same as in the pitch evaluation. (a) The duration reconstruction becomes more accurate when fine-tuning with limited accented speech data. (b) The overall performance of the three comparatively fine-tuned systems is ranked in a descending order: G2P-F0_Dur_AM-LS, G2P-AM-LS, and Char-AM-S. This observation is also consistent on Scottish and Australian accents. The results further confirm the contribution of the duration predictor to the duration reconstruction.

It is noted that the pitch and duration predictors highly rely on the small amount of accented speech training data.

*2) Subjective Evaluation:* We evaluate the speech quality in terms of naturalness by mean opinion score (MOS) [63] and

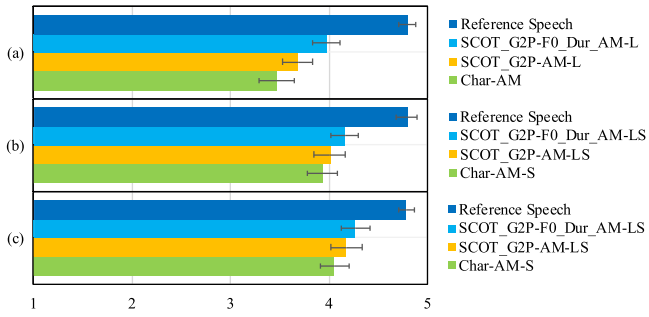


Fig. 8. MOS test results for naturalness of Char-AM, SCOT_G2P-AM-L, SCOT_G2P-F0_Dur_AM-L and Reference Speech on Scottish accent with 95% confidence intervals. (a) Not fine-tuned with accented speech data. (b) Fine-tuned with 50 utterances. (c) Fine-tuned with 300 utterances.

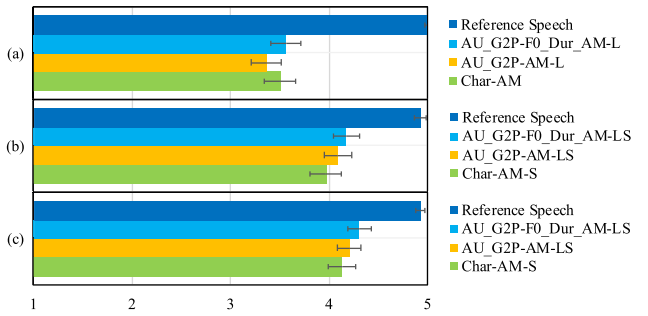


Fig. 9. MOS test results for naturalness of Char-AM, AU_G2P-AM-L, AU_G2P-F0_Dur_AM-L and Reference Speech on Australian accent with 95% confidence intervals. (a) Not fine-tuned with accented speech data. (b) Fine-tuned with 50 utterances. (c) Fine-tuned with 300 utterances.

further conduct XAB preference test and best-worst scaling (BWS) [64] test to assess the accent similarity.

*i) Speech Quality:* In the MOS test, the listeners are asked to rate the speech naturalness of the provided samples on a 5-point scale from 1 to 5. Each listener rates 24 samples in a single experiment, thus in a total of 144 ($24 \times 3$ (# of fine-tuning cases) $\times$ 2 (# of accents) $=$ 144) samples. We evaluate the three comparative TTS systems and reference speech under three different scenarios, (a) Not fine-tuned with accented speech data, (b) Fine-tuned with 50 accented speech utterances, (c) Fine-tuned with 300 accented speech utterances, on both Scottish and Australian accents. The results are presented in Fig. 8 and Fig. 9, respectively.

In Fig. 8, it is observed that SCOT_G2P-F0_Dur_AM-L achieves the highest score in all three cases, followed by SCOT_G2P-AM-L and Char-AM. This shows that phoneme input, pitch, and duration predictors help to generate higher quality and more natural speech. A similar conclusion can be drawn for Australian accent from Fig. 9. However, we note that in Fig. 9(a), the performance of AU_G2P-F0_Dur_AM-L is not obviously better than that of Char-AM, and even AU_G2P-AM-L obtains the slightly lower score than Char-AM. While in Fig. 8(a) the performance of SCOT_G2P-AM-L is much better than that of Char-AM. We suspect that the phoneme representation of the Australian accent is more distinctive than the Scottish accent when compared with the pre-training data of General American

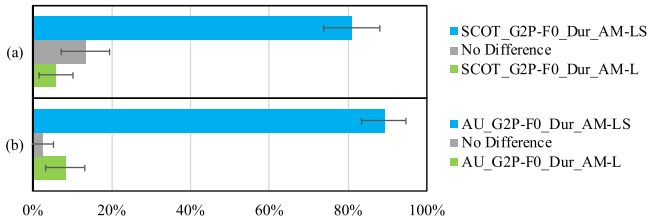


Fig. 10. XAB test results for accent similarity of G2P-F0_Dur_AM-L and G2P-F0_Dur_AM-LS on both Scottish and Australian accents with 95% confidence intervals. (a) Scottish accent. (b) Australian accent.

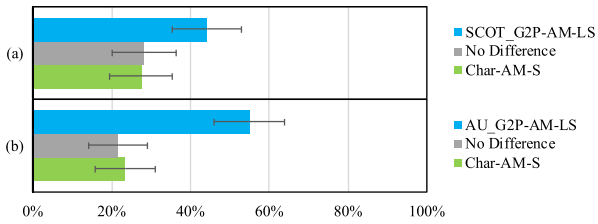


Fig. 11. XAB test results for accent similarity of Char-AM-S and G2P-AM-LS on both Scottish and Australian accents with 95% confidence intervals. (a) Scottish accent. (b) Australian accent.

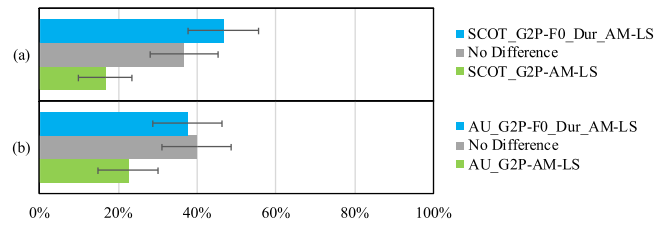


Fig. 12. XAB test results for accent similarity of G2P-AM-LS and G2PF0_Dur_AM-LS on both Scottish and Australian accents with 95% confidence intervals. (a) Scottish accent. (b) Australian accent.

English, Irish English, and British English Received Pronunciation. The unseen phonemes in Australian accent adversely affect the speech quality and naturalness to some extent. This may be also the reason that both G2P-AM-L in Fig. 8(a) achieve better performance than those two in Fig. 9(a).

*ii) Accent Similarity:* In the both XAB preference and BWS tests, all TTS systems with '-LS' denote that the TTS acoustic models are fine-tuned with 50 accented speech utterances.

In the XAB preference test, the listeners are asked to listen to a reference speech first and then select the more similar sample to the reference speech from two different samples according to the accent similarity. Each listener listens to 18 samples in a single experiment, thus in a total of 108 ($18 \times 3$ (# of comparison pairs) $\times$ 2 (# of accents) $=$ 108) samples. The results are shown in Figs. 10–12. In Fig. 10, it is apparent that the fine-tuned G2P-F0_Dur_AM-LS significantly outperforms the G2P-F0_Dur_AM-L. This shows that fine-tuning TTS acoustic model with limited accented speech data improves the prosodic rendering of generated accented speech. Fig. 11 shows that the generated speech from TTS acoustic model with correct phoneme representation input has higher accent similarity than that with character input. This demonstrates that the phoneme-based encoder extracts the prosody-related textual

TABLE VI
BWS TEST RESULTS FOR ACCENT SIMILARITY OF THE FOUR COMPARISON SYSTEMS ON BOTH SCOTTISH AND AUSTRALIAN ACCENTS

| Accent | System | Best (%) | Worst (%) |
|---|---|---|---|
| Scottish | SCOT_G2P-F0_Dur_AM-L | 7.5 | 78.33 |
| | Char-AM-S | 13.33 | 13.33 |
| | SCOT_G2P-AM-LS | 25.83 | 5.00 |
| | SCOT_G2P-F0_Dur_AM-LS | **53.33** | **3.33** |
| Australian | AU_G2P-F0_Dur_AM-L | 0.83 | 86.67 |
| | Char-AM-S | 15.00 | 6.67 |
| | AU_G2P-AM-LS | 36.67 | 4.17 |
| | AU_G2P-F0_Dur_AM-LS | **47.50** | **2.50** |

information well, compared with the character-based encoder. In Fig. 12, we are glad to see that after involving pitch and duration predictors into the TTS acoustic model, there is an obvious improvement in the perception of accent similarity in the generated speech. This strongly shows the effectiveness of the integrated pitch and duration predictors on accent rendering. All the reported conclusions are consistent on both Scottish and Australian accents.

In the BWS test, we provide 5 samples of the same content to the listeners. They are asked to listen to the reference speech first and then choose the most and the least similar samples to the reference speech from four different samples according to the accent similarity. Each listener listens to 30 samples in a single experiment, thus in a total of 60 ($30 \times 2$ (# of accents) $= 60$) samples. The results on Scottish and Australian accents are shown in Table VI. It is observed that the proposed G2P-F0_Dur_AM-LS system obtains the highest best score and the lowest worst score among the four comparative systems. We also note that the G2P-F0_Dur_AM-L obtains significantly the lowest best score and the highest worst score. This indicates the effectiveness of fine-tuning with limited accented speech data on accent rendering. From the overall trend, the system performance on accent similarity can be ranked in an ascending order: G2P-F0_Dur_AM-L, Char-AM-S, G2P-AM-LS, and G2P-F0_Dur_AM-LS. The obtained conclusions are the same as those of the objective evaluations.

## VI. CONCLUSION

We propose and validate an accented TTS framework by addressing both phonetic and prosodic variations of accent rendering. The study shows that the phonetic variation can be modeled by an accented front-end with a small accented phonetic lexicon. Meanwhile, prosodic variation can be modeled by an accented TTS acoustic model with explicit pitch and duration control when a limited amount of accented speech data is available. The study also reveals that the accented front-end also contributes to accented prosodic rendering. The key finding of this work is that it is possible to effectively model a target accent with a limited amount of accented lexical and speech data. In the future work, we will explore a joint training framework that includes an accented front-end and an accented acoustic model for accented TTS synthesis. In addition, we also plan to conduct experiments on other E2E TTS acoustic models to increase the generalization of our proposed method.

## REFERENCES


[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7962–7966.

[3] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[4] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," *Interspeech*, ISCA, 2017.

[5] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[6] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 6706–6713.

[7] Y. Ren et al., "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32.

[8] Y. Ren et al., "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Representations*, 2020.

[9] W. Ping et al., "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. Int. Conf. Learn. Representations*, 2018.

[10] M. Chen et al., "Multispeech: Multi-speaker text to speech with transformer," 2020, *arXiv:2006.04664*.

[11] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, vol. 31.

[12] M. Whitehill, S. Ma, D. McDuff, and Y. Song, "Multi-reference neural TTS stylization with adversarial cycle consistency," 2019, *arXiv:1910.11958*.

[13] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5734–5738.

[14] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *Proc. IEEE 12th Int. Symp. Chin. Spoken Lang. Process.*, 2021, pp. 1–5.

[15] M. J. Munro, "Foreign accent and speech intelligibility," *Phonology Second Lang. Acquisition*, vol. 5, pp. 193–218, 2008.

[16] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech Perception Linguistic Experience: Issues Cross-Lang. Res.*, vol. 92, pp. 233–277, 1995.

[17] C. T. Best et al., "The emergence of native-language phonological influences in infants: A perceptual assimilation model," *Develop. Speech Perception: Transition Speech Sounds Spoken Words*, vol. 167, no. 224, pp. 233–277, 1994.

[18] A. Behrman, "Segmental and prosodic approaches to accent management," *Amer. J. Speech- Lang. Pathol.*, vol. 23, no. 4, pp. 546–561, 2014.

[19] J. Jügler, F. Zimmerer, J. Trouvain, and B. Möbius, "The perceptual effect of L1 prosody transplantation on L2 speech: The case of French accented German," in *Proc. InterSpeech*, 2016, pp. 67–71.

[20] L. Loots and T. Niesler, "Automatic conversion between pronunciations of different English accents," *Speech Commun.*, vol. 53, no. 1, pp. 75–84, 2011.

[21] P. B. De Mareüil and B. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, no. 4, pp. 247–267, 2006.

[22] Q. Yan, S. Vaseghi, D. Rentzos, C.-H. Ho, and E. Turajlic, "Analysis of acoustic correlates of British, Australian and American accents," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2003, pp. 345–350.

[23] D. L. Bolinger, "A theory of pitch accent in english," *Word*, vol. 14, no. 2/3, pp. 109–149, 1958.

[24] S. V. Levi, "Acoustic correlates of lexical accent in Turkish," *J. Int. Phonetic Assoc.*, vol. 35, no. 1, pp. 73–97, 2005.

[25] S. Winters and M. G. O'Brien, "Perceived accentedness and intelligibility: The relative contributions of F0 and duration," *Speech Commun.*, vol. 55, no. 3, pp. 486–507, 2013.

[26] L. Rognoni and M. G. Busá, "Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating: An experiment using prosody transplantation," in *Proc. Int. Symp. Acquisition Second Lang. Speech, Concordia Work. Papers Appl. Linguistics*, 2014, vol. 5, pp. 547–560.

[27] A. Deri and K. Knight, "Grapheme-to-phoneme models for (almost) any language," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2016, pp. 399–408.
[28] M. Waseem and C. Sujatha, "Speech synthesis system for Indian accent using festvox," *Int. J. Sci. Eng. Technol. Res.*, vol. 03, pp. 6903–6911, 2014.
[29] S. Kayte, M. Mundada, and D. C. Kayte, "Speech synthesis system for Marathi accent using festvox," *Int. J. Comput. Appl.*, vol. 130, no. 6, pp. 38–42, 2015.
[30] B. Kolluru, V. Wan, J. Latorre, K. Yanagisawa, and M. J. Gales, "Generating multiple-accent pronunciations for TTS using joint sequence model interpolation," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014.
[31] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Accent group modeling for improved prosody in statistical parameteric speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6890–6894.
[32] B. Abeysinghe, J. James, C. I. Watson, and F. Marattukalam, "Visualising model training via vowel space for text-to-speech systems," ResearchSpace, Auckland, New Zealand, 2022.
[33] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable accented text-to-speech synthesis," 2022, *arXiv:2209.10804*.
[34] J. Melechovsky, A. Mehrish, B. Sisman, and D. Herremans, "Accented text-to-speech synthesis with a conditional variational autoencoder," 2022, *arXiv:2211.03316*.
[35] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6905–6909.
[36] Y. Yasuda and T. Toda, "Investigation of Japanese PnG BERT language model in text-to-speech synthesis for pitch accent language," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1319–1328, Oct. 2022.
[37] A. Parlikar, S. Sitaram, A. Wilkinson, and A. W. Black, "The festvox indic frontend for grapheme to phoneme conversion," in *Proc. WILDRE: Workshop Indian Lang. Data- Resour. Eval.*, 2016.
[38] J. Pan et al., "A unified sequence-to-sequence front-end model for mandarin text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6689–6693.
[39] S. Bansal, A. Mukherjee, S. Satpal, and R. Mehta, "On improving code mixed speech synthesis with mixlingual grapheme-to-phoneme model," in *Proc. InterSpeech*, 2020, pp. 2957–2961.
[40] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6199–6203.
[41] Y. Sevinj, N. Géza, and G.-T. Bálint, "Transformer based grapheme-to-phoneme conversion," in *Proc. InterSpeech*, 2019, pp. 2095–2099.
[42] L. Dong, Z.-Q. Guo, C.-H. Tan, Y.-J. Hu, Y. Jiang, and Z.-H. Ling, "Neural grapheme-to-phoneme conversion with pre-trained grapheme models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6202–6206.
[43] E. Engelhart, M. Elyasi, and G. Bharaj, "Grapheme-to-phoneme transformer model for transfer learning dialects," 2021, *arXiv:2104.04091*.
[44] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5545–5549.
[45] A. Gutkin, "Uniform multilingual multi-speaker acoustic model for statistical parametric speech synthesis of low-resourced languages," 2017.
[46] J.-H. Kim, S.-H. Lee, J.-H. Lee, H.-G. Jung, and S.-W. Lee, "GC-TTS: Few-shot speaker adaptation with geometric constraints," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2021, pp. 1172–1177.
[47] A. Sokolov, T. Rohlin, and A. Rastrow, "Neural machine translation for multilingual grapheme-to-phoneme conversion," 2020, *arXiv:2006.14194*.
[48] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
[49] Z. Zhang, Q. Tian, H. Lu, L.-H. Chen, and S. Liu, "Adadurian: Few-shot adaptation for neural text-to-speech with durian," 2020, *arXiv:2005.05642*.
[50] S. Fitt, "Morphological approaches for an English pronunciation lexicon," *Int. Speech Commun. Assoc.*, 2001.
[51] H. Zen et al., "Libritts: A corpus derived from librispeech for text-to-speech," *Interspeech*, ISCA, 2019.
[52] C. Veaux et al., "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," in *Proc. Univ. Edinburgh. Centre Speech Technol. Res.*, vol. 6, 2017, p. 15.
[53] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Proc. 5th ISCA Workshop Speech Synth.*, 2004.
[54] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," 2018, *arXiv:1808.10583*.
[55] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4879–4883.
[56] D. P. Kingma, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representation*, 2014.
[57] J. M. Joyce, "Kullback-leibler divergence," in *International Encyclopedia of Statistical Science*. New York, NY, USA: Springer, 2011, pp. 720–722.
[58] Q. Yan, S. Vaseghi, D. Rentzos, and C.-H. Ho, "Analysis and synthesis of formant spaces of British, Australian, and American accents," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 676–689, Feb. 2007.
[59] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1849–1863, 2020.
[60] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS training with frame and style reconstruction loss," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1806–1818, 2021.
[61] M. Müller, "Dynamic time warping," Information Retrieval Music Motion. Berlin, Germany: Springer, 2007, pp. 69–84.
[62] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7962–7966.
[63] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, 2016.
[64] J. A. Lee, G. Soutar, and J. Louviere, "The best–worst scaling approach: An alternative to Schwartz's values survey," *J. Pers. Assessment*, vol. 90, no. 4, pp. 335–347, 2008.