Decoding Knowledge Transfer for Neural Text-to-Speech Training

Rui Liu, *Member, IEEE*, Berrak Sisman, *Member, IEEE*, Guanglai Gao, Haizhou Li, *Fellow, IEEE*

***Abstract*—Neural end-to-end text-to-speech (TTS) is superior to conventional statistical methods in many ways. However, the exposure bias problem, that arises from the mismatch between the training and inference process in autoregressive models, remains an issue. It often leads to performance degradation in face of out-of-domain test data. To address this problem, we study a novel decoding knowledge transfer strategy, and propose a multi-teacher knowledge distillation (MT-KD) network for Tacotron2 TTS model. The idea is to pre-train two Tacotron2 TTS teacher models in teacher forcing and scheduled sampling modes, and transfer the pre-trained knowledge to a student model that performs free running decoding. We show that the MT-KD network provides an adequate platform for neural TTS training, where the student model learns to emulate the behaviors of the two teachers, at the same time, minimizing the mismatch between training and run-time inference. Experiments on both Chinese and English data show that MT-KD system consistently outperforms the competitive baselines in terms of naturalness, robustness and expressiveness for in-domain and out-of-domain test data. Furthermore, we show that knowledge distillation outperforms adversarial learning and data augmentation in addressing the exposure bias problem.**

***Index Terms*—End-to-end TTS, autoregressive model, exposure bias, knowledge transfer, knowledge distillation**

I. Introduction

WITH the advent of deep learning, text-to-speech (TTS) studies have seen significant progress. The neural TTS solutions [1], [2] are the examples. Unlike the conventional TTS pipeline [3]–[6], the neural solutions employ a network to learn the mapping directly from the <text, wav> pair [7]. For instance, Tacotron [1], Tacotron2 [2] and their variants [8]–[10] are based on an encoder-decoder architecture with an attention mechanism [11]. The recurrent neural network (RNN) [12] based decoder predicts the acoustic features frame-by-frame in an autoregressive manner [13], [14]. Furthermore, the studies of novel neural vocoders [2], [15], [16] also greatly improve the speech quality.

Despite much progress, run-time stability remains a challenge for neural TTS [17]–[20]. The autoregressive based TTS decoding often produces unpredictable outputs during run-time inference, that include deletion or repetition of words, incomplete utterance, and inappropriate prosody phrase breaks [21]–[24], especially for out-of-domain text input. It suffers from the exposure bias problem [25], [26] that arises from the mismatch between training and inference data, and adversely affects the autoregressive decoding process [13], [14]. Typically, in the training stage, the decoder predicts the current frame based on the natural speech frames in the previous steps in a *teacher forcing* (TF) mode. There is no doubt that the teacher forcing mode optimizes the decoder to produce samples close to the ground-truth. However, during run-time inference, the natural speech frames are replaced by predicted speech frames in a *free running* (FR) mode.

To address the exposure bias problem, there have been studies on data augmentation strategy, which uses predicted sequence during training. Bengio et al. [27] proposed to mix ground-truth natural and predicted sequences in training stage with a sampling strategy, i.e., *scheduled sampling* (SS), where the decoder takes one of the two kinds of sequences randomly as input. While the scheduled sampling strategy clearly shows improvements in terms of the robustness [28], it has an adverse effect on the audio quality as it ignores the temporal dependency of the acoustic sequence [25], [28]. More recently, generative adversarial network (GAN) [29] is studied, such as professor forcing [30], [31], which learns to generalize the model trained in teacher forcing model for free running decoding. With the generative adversarial network, knowledge from teacher forcing model is transferred to free running decoder via adversarial learning. All the studies point to the direction that increasing the exposure to the predicted samples during training is an effective way to remedy the exposure bias problem. The question is how to make full use of the knowledge from both ground-truth sequence and predicted sequence in the training stage, which is the focus of this paper.

Recently, knowledge distillation has attracted increasing attention, in which a student model is trained to emulate the behavior of a teacher model [32]. In other words, knowledge is transferred from the teacher model to the student model. In our

This paper was submitted on 11 June 2021 for review. The research by Rui Liu was funded by the High-level Talents Introduction Project of Inner Mongolia University, with Rui Liu as the Principal Investigator. The research by Rui Liu and Berrak Sisman was also funded by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion (SRG ISTD 2020 158) and SUTD AI Project (SGPAIRS1821) Discovery by AI - The Understanding and Synthesis of Expressive Speech by AI. The work by Haizhou Li was partly supported by the Agency of Science, Technology and Research, Singapore, through the National Robotics Program under Grant No. 192 25 00054, and by Programmatic Grant No. A18A2b0046 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain). (Corresponding author: Guanglai Gao)

Rui Liu is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China. He is also with the Department of Electrical and Computer Engineering, National University of Singapore, and Singapore University of Technology and Design (SUTD). (e-mail: liurui_imu@163.com).

Berrak Sisman is with Singapore University of Technology and Design (SUTD). (e-mail: berrak_sisman@sutd.edu.sg).

Guanglai Gao is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China. (e-mail: csggl@imu.edu.cn)

Haizhou Li is with School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China. He is also with University of Bremen, Faculty 3 Computer Science / Mathematics, Enrique-Schmidt-Str. 5 Cartesium, 28359 Bremen, Germany (e-mail: haizhouli@cuhk.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2022.3171974, IEEE/ACM Transactions on Audio, Speech, and Language Processing

previous work [33], we show that knowledge distillation can serve as an appropriate framework to transfer the knowledge of temporal dependency of speech to a free running decoder. In this paper, we formulate a novel decoding strategy based on knowledge distillation to address the exposure bias problem.

Inspired by the previous findings, we would like to incorporate teacher forcing, scheduled sampling, and free running decoding modes into a single training process to benefit from the best of all. This paper makes the following contributions: 1) We formulate a decoding knowledge transfer paradigm for mitigating the exposure bias problem, and discuss multiple possible implementations, including data augmentation, adversarial learning and knowledge distillation; 2) We propose a novel multi-teacher knowledge distillation (MT-KD) strategy to transfer knowledge from teacher forcing and scheduled sampling models to a free running decoder; 3) We show that knowledge distillation is more effective than adversarial learning and data augmentation in addressing the exposure bias problem. To the best of our knowledge, this is the first attempt to leverage the knowledge of multiple decoding strategies into the training of a single TTS network.

While this work shares a similar motivation with our previous work [33] in terms of teacher-student training, it is different in many ways. 1) We formulate a general paradigm for decoding knowledge transfer that addresses the exposure bias problem, and its multiple implementations, of which the teacher-forcing knowledge distillation in [33] is just a special case; and 2) We propose a novel multi-teacher knowledge distillation scheme for Tacotron TTS under the decoding knowledge transfer paradigm, that wasn't discussed in [33].

The rest of this paper is organized as follows. In Section II, we discuss the background to motivate our research. In Section III, we study the proposed multi-teacher knowledge distillation scheme for decoding knowledge transfer. We report the experiments in Section IV. Finally, Section VI concludes the study.

II. Background

We first review the Tacotron2 end-to-end TTS model [2], followed by a brief description of TTS training strategies and exposure bias. At last, we discuss the background of decoding knowledge transfer to set the stage for our study.

A. Tacotron2 TTS model

An encoder-decoder network learns to map an input sequence to an output sequence. Tacotron2 TTS system is one of the successful encoder-decoder network implementations, as illustrated in Figure 1. It consists of an encoder, attention-based decoder and two alternatives for waveform generation, that are described next in detail.

1) Encoder: The encoder converts the input text to feature representations [2]. It consists of a CNN [34] module armed with 3 convolution layers, and a bidirectional LSTM [35] module.

2) Decoder: The decoder takes the feature representations as input and predicts the hidden states [2], which are then converted to mel-spectrum features by a fully connected

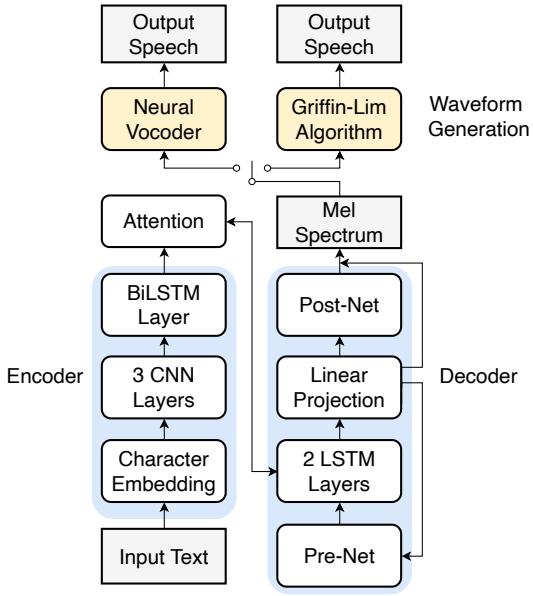

Fig. 1: Block diagram of Tacotron2 based reference baseline that has three modules, an encoder, an attention-based decoder, and two alternative methods for waveform generation.

operation [2]. It consists of a 2-layer pre-net, 2 LSTM layers, a linear projection layer, and a post-net of 5 convolution layers. The decoder is a standard autoregressive recurrent neural network that generates the mel-spectrum features and stop tokens frame by frame. The location-sensitive attention [36] is applied to learn the linguistic-acoustic alignment.

3) Waveform Generation: To synthesize the output audio from the mel-spectrum features, there are two common techniques. We may either use Griffin Lim [37] algorithm, or alternatively a WaveNet-based neural vocoder [7].

During training, the decoder predicts the current frame in a teacher forcing mode. However, during run-time inference, the decoder performs in a free running mode because the natural speech frames are unavailable, that leads to the exposure bias problem.

B. TTS decoding strategy

There have been studies to address the exposure bias problems in image generation [38], image captioning [39]–[41], text generation [25] and neural machine translation [42]. The exposure bias problem in TTS is fundamentally caused by the autoregressive decoding mismatch between training and inference [30].

In Figure 2, we summarize three decoding strategies in TTS. As shown in Figure 2 (a), with a teacher forcing mode, the decoder predicts current frame y' using its previous natural speech frames $y_{<t}$ as input [14]. The probability distribution $p(y', x)$ can be formulated as follows,

$$p(y', x) = \sum_{t=1}^{T'} p(y'_t | y_{<t}, x) \tag{1}$$

As the training process only exposes the model to natural speech data, it optimizes the model distribution to be close

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2022.3171974, IEEE/ACM Transactions on Audio, Speech, and Language Processing

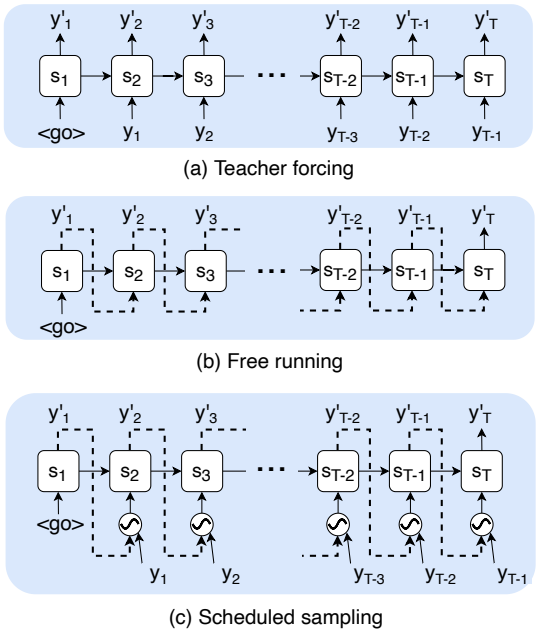

Fig. 2: Three decoding strategies of autoregressive model. (a) teacher forcing, (b) free running, and (c) scheduled sampling, ∿ represents the sampling function.

to that of natural speech data. However, during run-time inference, a decoder always runs in a free running mode. In other words, the decoder predicts a frame y' using its previously predicted frames $y'_{<t}$ instead of natural speech frames as the history, that is illustrated in Figure 2 (b). The probability distribution $p(y', x)$ can be expressed as follows,

$$p(y', x) = \sum_{t=1}^{T'} p(y'_t | y'_{<t}, x) \tag{2}$$

As a result, the predicted speech frames at training and inference are drawn from different distributions, i.e., natural speech data vs model predicted data, resulting in accumulated prediction errors.

In recent neural TTS studies [1], [43], a scheduled sampling strategy [27] is introduced during training as a way to mitigate the exposure bias. As shown in Figure 2 (c), at each time step, the training process decides whether a natural speech frame or a predicted frame is to be added to the prediction history.

$$p(y', x) = \sum_{t=1}^{T'} p(y'_t | \mathcal{S}(y_{<t}, y'_{<t}), x) \tag{3}$$

where $\mathcal{S}(\cdot)$ means the sampling function, which is represented by ∿ in Figure 2 (c). Scheduled sampling is shown to be effective in learning robust model distribution [26] as it exposes the model to both natural speech data and predicted data.

The scheduled sampling strategy seeks to address the exposure bias problem by augmenting teacher-forcing data with free-running data. While it mitigates the exposure bias problem to some extent, it leads to other issues such as speech discontinuity because of the misalignment between the real data and the predicted data. In other words, it doesn't consider the temporal dependency of the acoustic sequence [25], [28], [44]. The ignorance of temporal dependency has an adverse impact on speech quality [44]. Furthermore, as the inference stage is in a complete free-running mode, the scheduled sampling strategy doesn't fully solve the exposure bias problem.

C. Decoding knowledge transfer

Besides the data augmentation strategy, another school of thought to address the exposure bias problem is to transfer the decoding knowledge from one model to another, instead of solely training the model from data. The studies of knowledge transfer can be generally grouped into two categories, adversarial learning [29] and knowledge distillation [32].

Adversarial learning, such as GAN, has been used in many sequential training tasks, for example, text classification [45], machine translation [46], and speech recognition [47]. It is also used to learn a domain-invariant representation to improve model generalization in target domains [48]. In practice, one can employ a pre-trained discriminator to distinguish between the hidden state sequences generated by two models in different domains [30]. With such adversarial learning, the discriminator seeks to minimize the difference of behaviors between the two models so as to transfer knowledge from one another.

Knowledge distillation is another technique for knowledge transfer, usually from a high-performing teacher network into a simple student network [32]. To this end, the probability distribution derived by a teacher network is regarded as the 'soft target' to help the student network to learn the behavior of the teacher network. Knowledge distillation has been successfully applied to natural language processing [49], [50], computer vision [51], [52] and speech processing [53], [54]. Recently, knowledge distillation with multiple teachers was studied to integrate multi-level knowledge [55], to perform ensemble learning for model compression [56], [57].

In the context of TTS synthesis, knowledge transfer has been studied as one of the ways to transfer the desired decoding knowledge to the run-time decoder. The professor forcing strategy [44] is one such example, which seeks to improve the generalization ability of TTS model by jointly training a discriminator in a GAN architecture so as to minimize the hidden states between two generators running in teacher forcing (or scheduled sampling) mode and free running mode respectively. Unfortunately, the implementation of the adversarial learning process is not straightforward [58]. In the professor forcing strategy, the two generators act as peers to each other without the distinctive role of 'professor' or 'teacher'. Furthermore, the quality of the pre-trained discriminator has an impact on the overall performance. Nonetheless, the study of professor forcing strategy suggests that teacher forcing and scheduled sampling decoding knowledge each has it own merits. The question is how the free running decoding can benefit from the best of both.

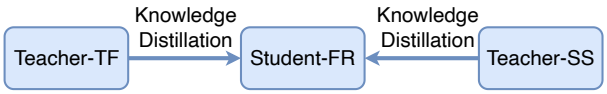

Fig. 3: The proposed multi-teacher knowledge distillation (MT-KD) training strategy, that consists of two teacher models and one student model.

A knowledge distillation strategy [33] was studied to address the exposure bias problem differently from the scheduled sampling strategy. In [33], the student model is trained in a free-running mode, with the knowledge transferred from a teacher model pre-trained in teacher forcing mode. We consider that knowledge distillation technique provides an adequate platform for transferring the decoding knowledge from one model to another not only because it benefits from well pre-defined teacher models, but also from models of multiple desired behaviors. As the student model is trained in a free-running mode, it is inherently consistent with the inference process. In this paper, we expand the idea of knowledge distillation in [33] to a decoding knowledge transfer paradigm.

We note that the teacher-student transfer learning as a knowledge distillation technique clearly defines the role of teacher and student models, while the professor forcing training as an adversarial learning technique doesn't. We consider that the former is better motivated than the latter as far as decoding knowledge transfer is concerned. We will study a new way to take full advantage of both teacher forcing and scheduled sampling decoding knowledge for the free running run-time decoder.

III. Multi-Teacher Knowledge Distillation for Decoding Knowledge Transfer

We propose a multi-teacher knowledge distillation (MT-KD) training strategy in Figure 3, that employs two teacher models to guide the training of a student model, for a neural TTS system. All teacher and student models share a similar network architecture as *Tacotron2* [2] in Section II-A.

The training of MT-KD is conducted in two phases. First, we pre-train two teacher models, one in teacher-forcing mode (*Teacher-TF*) and another in scheduled sampling mode (*Teacher-SS*), following the *Tacotron2* training protocol. We then train the *Student-FR* model in a similar way, except that its decoder is trained in a free running mode, and jointly supervised by two pre-trained decoders, and the target speech. At run-time, the *Student-FR* model predicts the speech frames in a free running mode that is consistent with its training mode. Next we discuss the training and inference process in detail.

A. Phase I: Multi-teacher pre-training

1) Teacher-TF pre-training: As shown in Figure 4, *Teacher-TF* consists of an Encoder ($\mathrm{Enc_{TF}}$) and an attention-based Decoder ($\mathrm{Dec_{TF}}$), where $\mathrm{Dec_{TF}}$ is trained in a teacher forcing mode, which takes the previous natural speech frames as the input to predict the current speech frame.

Next we briefly recap the training process. Given an input character sequence $x = (x_1, x_2, ..., x_T)$ and its target mel-spectrum features $y = (y_1, y_2, ..., y_{T'})$, the text encoder $\mathrm{Enc_{TF}}$ reads x and outputs a hidden feature h_t at each step t ($t \in [1, T]$):

$$h_t = \mathrm{Enc_{TF}}(x_t) \tag{4}$$

The decoder $\mathrm{Dec_{TF}}$ takes the previous frames $y_1, ..., y_{t-1}$ from the target natural speech as input to output a new hidden state s_t at time step t, as formulated next,

$$s_t = \mathrm{Dec_{TF}}(s_{t-1}, y_{t-1}, \sigma(h_t)) \tag{5}$$

where y_{t-1} represents the ground truth acoustic feature, and $\sigma(\cdot)$ represents a function to calculate the context vector by using location-sensitive attention mechanism [36]. After this, a fully connected layer $g(\cdot)$ generates the mel-spectrum features $\hat{y}_t$ from the hidden states s_t,

$$\hat{y}_t = g(s_t) \tag{6}$$

The *Teacher-TF* model adopts the feature loss function $Loss_f$ to minimize the frame-level distance between the generated speech and the natural speech,

$$Loss_f = \sum_{t=1}^{T'} L_2(\hat{y}_t, y_t) \tag{7}$$

As the *Teacher-TF* model is trained on natural speech frame sequence as the input of decoder, we expect that the *Teacher-TF* model reflects the true distribution of the natural speech data, and capture the temporal dynamic of speech signal.

2) Teacher-SS pre-training: As shown in Figure 4, the *Teacher-SS* model also consists of an Encoder ($\mathrm{Enc_{SS}}$) and a Decoder ($\mathrm{Dec_{SS}}$). The *Teacher-SS* pretraining is performed in a similar way to the *Teacher-TF* except that its $\mathrm{Dec_{SS}}$ is trained in a scheduled sampling mode.

The $\mathrm{Enc_{SS}}$ performs the same task as $\mathrm{Enc_{TF}}$. However, the $\mathrm{Dec_{SS}}$ works differently from the $\mathrm{Dec_{TF}}$, which randomly takes the previous predicted frame $\hat{y}_{t-1}$ or the natural speech frame y_{t-1} as the input [27]. The decoding process of the $\mathrm{Dec_{SS}}$ is defined as:

$$s_t = \mathrm{Dec_{SS}}(s_{t-1}, \mathcal{S}(\hat{y}_{t-1}, y_{t-1}), \sigma(h_t)) \tag{8}$$

where $\mathcal{S}(\cdot)$ represents a sampling function as mentioned in Eq. 3. At last, we convert the hidden states s to the output sequence y and calculate the loss function $Loss_f$.

As the *Teacher-SS* model is trained on both natural and predicted speech frames, it has the adequate exposure to the distributions of both type of data, that reduces the mismatch between training and free running inference.

B. Phase II: Student training

The *Student-FR* model has the same network architecture as the teacher models, which includes $\mathrm{Enc_{FR}}$ and $\mathrm{Dec_{FR}}$. On the right panel of Figure 4, we illustrate the schematic diagram of the training process. To benefit from the pre-trained models, we use $\mathrm{Enc_{TF}}$ or $\mathrm{Enc_{SS}}$ to initialize $\mathrm{Enc_{FR}}$,

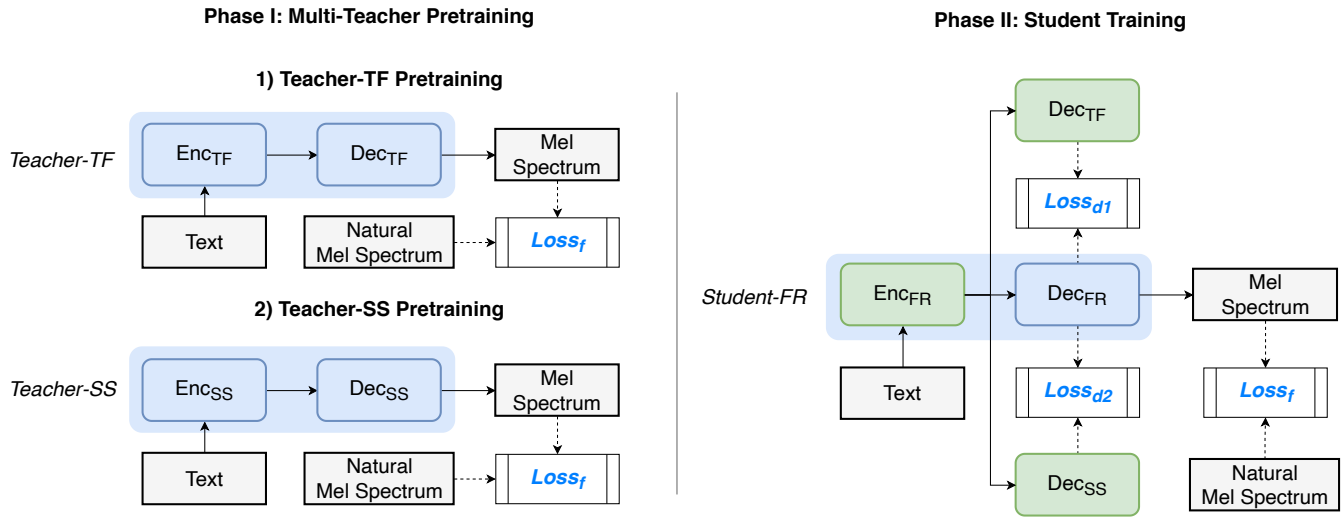

Fig. 4: An illustration of our multi-teacher knowledge distillation (MT-KD) scheme for a neural TTS system. The parameters in the blue boxes are initialized with random values, while those in green boxes are initialized by taking the pretrained parameters from Phase I.

which generates the linguistic encoding from input text. In a pilot experiment, we found no difference between the two initialization methods in terms of naturalness. Therefore, we adopt $\mathrm{Enc_{TF}}$ to initialize the student encoder $\mathrm{Enc_{FR}}$ in future experiments.

Next, the student encoder $\mathrm{Enc_{FR}}$ takes the given input text $x = (x_1, x_2, ..., x_T)$, and outputs linguistic encoding $h = (h_1, h_2, ..., h_T)$:

$$h = \mathrm{Enc_{FR}}(x) \tag{9}$$

We train the *Student-FR* model following the *Tacotron2* training process. During training, the teacher decoders and student decoder take the same encoder output sequence as the input, and generate their own hidden states respectively. We apply two distillation loss functions, $Loss_{d1}$ and $Loss_{d2}$, to supervise the hidden states derived by the student decoder to be close to those of the teacher models. At the same time, we adopt the feature loss function $Loss_f$ to ensure that the predicted speech is close to the reference natural speech. The decoder is trained in a free running mode that is consistent with the decoding process during run-time inference. We next formulate the training of $\mathrm{Dec_{FR}}$.

- The pre-trained $\mathrm{Dec_{TF}}$ takes the previous hidden states $s_{\mathrm{TF}_{t-1}}$, natural speech frame y_{t-1} and the attention score $\sigma(h_t)$ as input, and outputs the hidden state s_{TF_t} at each time step t as the Eq. 5:

$$s_{\mathrm{TF}_t} = \mathrm{Dec_{TF}}(s_{\mathrm{TF}_{t-1}}, y_{t-1}, \sigma(h_t)) \tag{10}$$

- The pre-trained $\mathrm{Dec_{SS}}$ outputs another hidden state s_{SS_t} at each time step t by following Eq. 8:

$$s_{\mathrm{SS}_t} = \mathrm{Dec_{SS}}(s_{\mathrm{SS}_{t-1}}, \mathcal{S}(\hat{y}_{t-1}, y_{t-1}), \sigma(h_t)) \tag{11}$$

- At the same time, the student decoder $\mathrm{Dec_{FR}}$ takes the previous hidden states $s_{\mathrm{FR}_{t-1}}$, estimated speech frame $\hat{y}_{t-1}$ and the attention score $\sigma(h_t)$ as input, and predicts the hidden state s_{FR_t} at each step t:

$$s_{\mathrm{FR}_t} = \mathrm{Dec_{FR}}(s_{\mathrm{FR}_{t-1}}, \hat{y}_{t-1}, \sigma(h_t)) \tag{12}$$

Finally, we follow Eq. 6 to generate the output speech frame $\hat{y}_t$:

$$\hat{y}_t = g(s_{\mathrm{FR}_t}) \tag{13}$$

The training is supervised by three three objective functions, $Loss_f$, $Loss_{d1}$ and $Loss_{d2}$. $Loss_f$ is the speech generation loss between the predicted speech $\hat{y}_t$ and the reference speech y_t,

$$Loss_f = \sum_{t=1}^{T'} L_2(\hat{y}_t, y_t) \tag{14}$$

The two distillation loss $Loss_{d1}$ and $Loss_{d2}$ ensure that the hidden states of $\mathrm{Dec_{FR}}$ are as close to $\mathrm{Dec_{TF}}$ and $\mathrm{Dec_{SS}}$ as possible.

$$Loss_{d1} = \frac{1}{T'} \sum_{t=1}^{T'} |s_{\mathrm{TF}} - s_{\mathrm{FR}}|^2 \tag{15}$$

$$Loss_{d2} = \frac{1}{T'} \sum_{t=1}^{T'} |s_{\mathrm{SS}} - s_{\mathrm{FR}}|^2 \tag{16}$$

We formulate a total loss for the training of *Student-FR* as follows,

$$Loss_{total} = Loss_f + \lambda \cdot Loss_{d1} + (1 - \lambda) \cdot Loss_{d2} \tag{17}$$

where λ is a scaling factor between the two distillation loss

Algorithm 1: The pseudocode of the MT-KD training.

Input:
Training set: $D = \{x, y\}$
x: character sequence
y: mel-spectrum feature sequence
Output:
Θ: TTS model including Enc and Dec
Begin
>> Phase-I: Multi-Teacher Pre-training
1: Initialize *Teacher-TF* model: Θ_{TF} ($\mathrm{Enc_{TF}}$ and $\mathrm{Dec_{TF}}$)
2: **for** each iteration **do**
$h = \mathrm{Enc_{TF}}(x)$
$s_{\mathrm{TF}} = \mathrm{Dec_{TF}}(h)$
$\hat{y} = g(s_{\mathrm{TF}})$
3: update Θ_{TF} with $Loss_f$:
$\Theta_{\mathrm{TF}} \leftarrow \bigtriangledown_{\Theta_{\mathrm{TF}}}(\mathrm{Loss_f(y, \hat{y})})$
4: return Θ_{TF} ($\mathrm{Enc_{TF}}$ and $\mathrm{Dec_{TF}}$)
5: Initialize *Teacher-SS* model: Θ_{SS} ($\mathrm{Enc_{SS}}$ and $\mathrm{Dec_{SS}}$)
6: **for** each iteration **do**
$h = \mathrm{Enc_{SS}}(x)$
$s_{\mathrm{SS}} = \mathrm{Dec_{SS}}(h)$
$\hat{y} = g(s_{\mathrm{SS}})$
7: update Θ_{SS} with $Loss_f$:
$\Theta_{\mathrm{SS}} \leftarrow \bigtriangledown_{\Theta_{\mathrm{SS}}}(\mathrm{Loss_f(y, \hat{y})})$
8: return Θ_{SS} ($\mathrm{Enc_{SS}}$ and $\mathrm{Dec_{SS}}$)
>> Phase-II: Student Training
1: Initialize *Student-FR* model Θ_{FR}:
$\mathrm{Enc_{FR}} \leftarrow \mathrm{Enc_{TF}}$ or $\mathrm{Enc_{SS}}$
2: Load pretrained teacher decoders $\mathrm{Dec_{TF}}$ and $\mathrm{Dec_{SS}}$:
$\mathrm{Dec_{TF}} \leftarrow \mathrm{Dec_{TF}}$
$\mathrm{Dec_{SS}} \leftarrow \mathrm{Dec_{SS}}$
3: **for** each iteration **do**
$h = \mathrm{Enc_{FR}}(x)$
$s_{\mathrm{FR}} = \mathrm{Dec_{FR}}(h)$
$s_{\mathrm{TF}} = \mathrm{Dec_{TF}}(h)$
$s_{\mathrm{SS}} = \mathrm{Dec_{SS}}(h)$
$\hat{y} = g(s_{\mathrm{FR}})$
4: update Θ_{FR} with total Loss:
$\Theta_{\mathrm{FR}} \leftarrow \bigtriangledown_{\Theta_{\mathrm{FR}}}(\mathrm{Loss_f(y, \hat{y})} + \lambda \cdot \mathrm{Loss_{d1}(s_{TF}, s_{FR})}$
$+(1-\lambda) \cdot Loss_{d2}(s_{\mathrm{SS}}, s_{\mathrm{FR}}))$
5: return Θ_{FR} ($\mathrm{Enc_{FR}}$ and $\mathrm{Dec_{FR}}$)
End

terms. Algorithm 1 describes the complete training process of the proposed MT-KD.

C. Run-time inference

At run-time, only *Student-FR* model is involved, where we use $\mathrm{Enc_{FR}}$ to process input text, and $\mathrm{Dec_{FR}}$ to generate an acoustic feature sequence. With multi-teacher knowledge distillation, we expect that *Student-FR* to generate natural, robust and expressive synthesized speech. We will validate the performance of the MT-KD training scheme in Section IV.

D. Decoding knowledge transfer paradigm

It is apparent that, if we remove the *Teacher-SS* from MT-KD, MT-KD is reduced to a 1-teacher knowledge distillation model, that is referred to as the teacher forcing knowledge distillation or TF-KD, as in our previous work [33]. By removing the *Teacher-TF* from MT-KD, we arrive at another 1-teacher knowledge distillation model, that is referred to as a scheduled sampling knowledge distillation model or SS-KD. In this paper, we will implement TF-KD and SS-KD together with MT-KD in a comparative study, where SS-KD and MT-KD are studied for the first time, while TF-KD is a re-implementation of [33]. The comparison of three different teacher-student training schemes are summarized in Table I for ease of reference. The TF-KD, SS-KD, and MT-KD systems are under the same decoding knowledge transfer paradigm based on knowledge distillation.

TABLE I: Three knowledge distillation schemes for decoding knowledge transfer. (TF: Teacher forcing; SS: Scheduled sampling; FR: Free running)

System	Teacher		Student
	TF	SS	FR
TF-KD [33]	yes	no	yes
SS-KD ***(new)***	no	yes	yes
MT-KD ***(new)***	yes	yes	yes

IV. Experiments

We conduct a series of experiments to benchmark the proposed multi-teacher knowledge distillation against other competitive training schemes under the same *Tacotron2* framework. Through comparative study, we would like to observe the individual contributions by *multi-teacher* and *knowledge distillation*.

A. Experimental data

We report the experiments on a Chinese and an English TTS dataset that are publicly available. The Chinese dataset, denoted as CSMSC[1], includes a total of 12 hours of standard Mandarin speech in 10,000 sentences by a native female speaker. The speech data are sampled at 48 kHz and encoded at 16 bits. The English dataset, denoted as LJSpeech[2] consists of 13,100 short clips with a total of nearly 24 hours of speech from one single speaker. The speech data are sampled at 22.05 kHz and encoded at 16 bits.

We partition all TTS corpora into training, validation, and test set at a ratio of 8:1:1. The training set is used for the training of all models, including pre-trained models. The validation set is used to regulate the early stopping scheme [59] to avoid overfitting to the training set. The test set serves as the in-domain evaluation data. We further select an additional set of data to serve as the out-of-domain evaluation data. We will discuss the overall evaluation data next.

We design two subsets, in-domain and out-of-domain, to form an evaluation dataset. The in-domain subset consists of 200 utterances from both Chinese and English test sets, resulting in a total of 400 utterances.

[1]https://www.data-baker.com/open_source.html
[2]https://keithito.com/LJ-Speech-Dataset/

The out-of-domain subset is a more challenging test set. We select 500 Chinese utterances from the Blizzard Challenge 2019 Chinese dataset [60]. These utterances have rich text content, including long sentences (180 characters on average), digital sequence and abbreviation, etc. We further collect 150 English utterances from FastSpeech [61] and ParaNet [62]. These utterances also include single letters, abbreviations, spellings, repeated numbers and long sentences (128 characters on average).

As the out-of-domain subset is not well covered during training, we expect that its TTS outputs have a lower intelligibility than those of the in-domain set. We use the in-domain set to conduct the objective and naturalness evaluations, while using both in-domain and out-of-domain set to compare the system robustness (intelligibility) and expressiveness.

B. Comparative study

We implement seven training schemes with the *Tacotron2* model in a comparative study. TF and SS are two *Tacotron2* models trained in teacher-forcing and scheduled sampling mode. We follow the TF-GAN and SS-GAN network architecture and training protocol in [44]. TF-GAN and SS-GAN are collectively referred to as adversarial learning systems, while TF-KD, SS-KD, and MT-KD are three teacher-student training schemes, collectively referred to as knowledge distillation systems.

- TF [2]: *Tacotron2* TTS model trained in the teacher forcing mode, which involves neither data augmentation, distillation, nor adversarial learning.
- SS [27]: *Tacotron2* TTS model trained in the scheduled sampling mode for data augmentation.
- TF-GAN [44]: *Tacotron2* TTS model with GAN-based training, which learns to minimize the difference between teacher forcing and free running decoding by adversarial learning.
- SS-GAN [44]: *Tacotron2* TTS model with GAN-based training, which learns to minimize the difference between scheduled sampling and free running decoding by adversarial learning.
- TF-KD [33]: *Tacotron2* TTS model with 1-teacher knowledge distillation, i.e., *Teacher-TF* only, that was previously reported in [33]. TF-KD seeks to transfer teacher forcing knowledge to free running decoder.
- SS-KD ***(new)***: *Tacotron2* TTS model with knowledge distillation of 1-teacher, i.e., *Teacher-SS* only. SS-KD seeks to transfer scheduled sampling knowledge to free running decoder.
- MT-KD ***(new)***: *Tacotron2* TTS model with the proposed multi-teacher knowledge distillation scheme.

We will compare the three knowledge distillation systems, i.e., TF-KD, SS-KD, and MT-KD, with the adversarial learning systems, i.e., TF-GAN and SS-GAN. We will also compare the 2-teacher MT-KD system with its 1-teacher counterparts, i.e., TF-KD, and SS-KD systems.

While the systems are trained with different training schemes, they perform decoding in a free running mode at run-time. We use Griffin-Lim algorithm [37] and Parallel-WaveGAN vocoder [63] for waveform generation. It is known that Griffin-Lim is a widely used waveform reconstruction algorithm, that compromises speech quality for rapid turn-around, while Parallel-WaveGAN is a real-time and small-footprint neural vocoder, that requires more computation and produces high speech quality in general. We will conduct comprehensive experiments with both waveform generation techniques. The speech samples are available at the demo site[3].

C. Experimental setup

The Chinese text is first converted to *pinyin*[4] sequence with tones, that share the character set with English text. We collectively call the Chinese and English text as character sequence, that is taken by the encoder as input. For both languages, the decoder takes a sequence of 256-dimensional encoder outputs as input and predicts a sequence of 80-channel mel-spectrum features. A mel-spectrum is extracted with 12.5 ms frame shift and 50 ms frame length. We follow the parameter settings in [5] to normalize all mel-spectrum features. The reduction factor is set to 2.

In TF and SS experiments, we train all *Tacotron2* models for 150 k steps. In TF-GAN and SS-GAN experiments, we follow the generator and discriminator architecture in [30], and the training method in [44]. In TF-KD, SS-KD and MT-KD experiments, all teacher models and student models are trained with 150 k steps.

In a preliminary study, we evaluate the effect of the linear decay [27] based scheduled sampling strategy in model training. We find that the quality and clarity of speech deteriorate significantly when we linearly decay the sampling probability from 1 to 0 [6]. When the sampling probability is reduced to 0, the training is reduced to a free-running mode. While the free-running mode matches the inference process, it doesn't guarantee that the model learns to predict the real data well. Our finding corroborates that of [44]. Therefore, we follow [44] and employ a scheduled sampling strategy with a linear decaying probability from 1 to 0.5 in the first 50 k steps, for SS, SS-GAN, SS-KD and MT-KD systems. Hyper-parameter λ in Eq. 17 is empirically set to 0.4.

All models are trained with a batch size of 32 and the Adam optimizer with β_1 = 0.9, β_2 = 0.999. The learning rate is exponentially decayed from 10^{-3} to 10^{-5} after 50k steps.

[3] https://ttslr.github.io/MT-KD/

[4] The standard romanization of Chinese.

[5] https://github.com/TensorSpeech/TensorFlowTTS/blob/master/preprocess/baker_preprocess.yaml and https://github.com/TensorSpeech/TensorFlowTTS/blob/master/preprocess/ljspeech_preprocess.yaml

[6] Note that this finding was based on the linear decay strategy and 150 K training steps. In our preliminary experiments, using some better decay strategies (such as exponential decay in the original paper of scheduled sampling) and training more steps (such as 600 k steps), we could further improve the decoding performance after the sampling probability drops to 0. However, It is not the focus of this paper to study how to improve the training strategy of SS during training stage. The experimental part of this paper mainly verifies the effectiveness of knowledge distillation in solving the exposure bias or decoding mode mismatch problem in SS while ensuring the fairness of the experiment.

TABLE II: A summary of objective and subjective evaluation experiments in terms of mel-spectrum distortion (MCD) and Root Mean Squared Error (RMSE), Mean Opinion Score (MOS), and Best Worst Scaling (BWS) over seven comparative systems. (GL: Griffin-Lim Algorithm; PW: Parallel-WaveGAN Vocoder)

System	Language	Objective Evaluation		MOS Evaluation		BWS Evaluation	
		MCD [dB]	RMSE [Hz]	GL	PW	Best (%)	Worst (%)
TF	English	8.48	20.35	3.68 ± 0.02	3.91 ± 0.03	0	66
	Chinese	8.17	20.28	3.61 ± 0.02	4.40 ± 0.05	0	80
SS*	English	8.32	20.22	3.70 ± 0.04	3.95 ± 0.02	0	22
	Chinese	7.99	20.12	3.62 ± 0.04	4.43 ± 0.03	0	14
TF-GAN†	English	8.33	20.17	3.69 ± 0.04	3.93 ± 0.03	4	5
	Chinese	7.98	20.09	3.70 ± 0.05	3.97 ± 0.02	8	6
SS-GAN†	English	8.31	20.18	3.70 ± 0.03	3.95 ± 0.05	0	7
	Chinese	7.92	20.15	3.71 ± 0.02	3.99 ± 0.03	0	0
TF-KD#	English	8.29	20.09	3.71 ± 0.01	4.00 ± 0.03	8	0
	Chinese	7.83	20.03	3.78 ± 0.02	4.02 ± 0.01	6	0
SS-KD#	English	8.27	20.10	3.79 ± 0.02	4.03 ± 0.04	6	0
	Chinese	7.85	20.01	3.80 ± 0.07	4.06 ± 0.05	0	0
MT-KD#	English	**8.13**	**19.98**	**3.82 ± 0.04**	**4.09 ± 0.03**	**82**	**0**
	Chinese	**7.61**	**19.85**	**3.85 ± 0.02**	**4.12 ± 0.01**	**86**	**0**
Ground Truth	English	NA	NA	4.39 ± 0.02		NA	NA
	Chinese	NA	NA	4.43 ± 0.03		NA	NA

(*: data augmentation; †: adversarial learning; #: knowledge distillation; both adversarial learning and knowledge distillation systems perform decoding knowledge transfer.)

D. Objective evaluation

We use Mel Cepstral Distortion (MCD) [64] and Root Mean Squared Error (RMSE) [4] as the objective evaluation metrics.

As the duration of the synthesized speech is usually different from that of the reference speech, we apply dynamic time warping (DTW) algorithm [65] to obtain a frame-level alignment between the two to facilitate MCD and RMSE calculation. We calculate a MCD between a reference speech and a synthesized speech of T frames as follows,

$$\text{MCD} = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\mathcal{N}}\sqrt{\sum_{k=1}^{\mathcal{N}}(y_{t,k}-\hat{y}_{t,k})^2}\right) \tag{18}$$

where $\mathcal{N}$ represents the dimension of the mel-spectrum, $y_{t,k}$ denotes the k^{th} mel-spectrum component of t^{th} frame for the reference speech, and $\hat{y}_{t,k}$ for that of the synthesized speech. Similarly, we calculate a RMSE between a reference speech and a synthesized speech of T frames as follows,

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\text{F0}_\text{t} - \widehat{\text{F0}}_t\right)^2} \tag{19}$$

where F0_t and $\widehat{\text{F0}}_t$ denote the reference and predicted F0 at t^{th} frame. Smaller value for both metrics indicates lower distortion, thus better performance.

1.1) Knowledge transfer vs data augmentation

The MCD and RMSE results are reported in the third and fourth columns of Table II. We observe that all models obtain lower MCD and RMSE than TF model, which corroborates the prior studies [27], [33], [44]. In other words, both data augmentation and knowledge transfer methods are effective in addressing the exposure bias problem. We note that knowledge transfer systems, including both knowledge distillation and adversarial learning systems, consistently outperform SS system, a data augmentation solution.

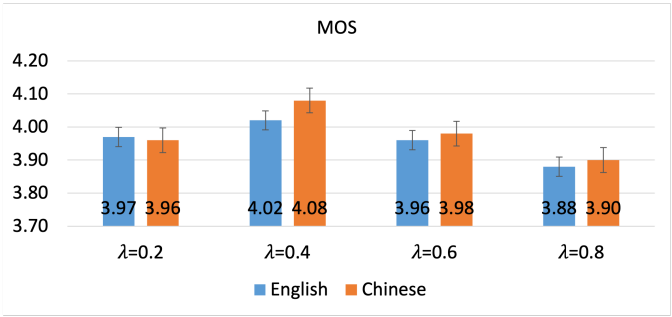

Fig. 5: Comparison of various trade-off parameter λ 0.2, 0.4, 0.6, 0.8 of the proposed MT-KD system in terms of mean opinion scores (MOS), with 95% confidence intervals computed from the t-test.

1.2) Knowledge distillation vs adversarial learning

We also observe that knowledge distillation systems consistently outperform their adversarial learning counterparts, i.e., TF-KD vs TF-GAN, and SS-KD vs SS-GAN. For English, TF-KD and TF-GAN achieve MCD of 8.29 dB and 8.33 dB respectively. For Chinese, they achieve 7.83 dB and 7.98 dB respectively. Similarly, TF-KD also achieves lower RMSE values than TF-GAN for both languages. We observe the same trend between SS-KD and SS-GAN.

These results suggest that knowledge distillation allows for the explicit transfer of the knowledge from a well-defined teacher to a student model. Adversarial learning is

TABLE III: The results of expressiveness evaluation in a listening experiment in terms of sentence-level Break Error Rate (BER %) for both in-domain and out-of-domain utterances.

System	Language	BER (%)	
		In-Domain	Out-of-Domain
TF	English	30	41
	Chinese	21	33
SS*	English	22	38
	Chinese	19	30
TF-GAN†	English	15	30
	Chinese.	14	29
SS-GAN†	English	14	27
	Chinese	12	25
TF-KD#	English	10	20
	Chinese	8	19
SS-KD#	English	9	18
	Chinese	6	12
MT-KD#	English	**2**	**8**
	Chinese	**0**	**5**

(*: data augmentation; †: adversarial learning; #: knowledge distillation; both adversarial learning and knowledge distillation systems perform decoding knowledge transfer.

less effective than knowledge distillation in this study.

1.3) 2-teacher KD vs 1-teacher KD

MT-KD, TF-KD, and SS-KD systems share the same network architecture and knowledge distillation strategy. The difference between MT-KD and TF-KD/SS-KD lies in the fact that MT-KD employs two teachers, while TF-KD and SS-KD only employ one of the two teachers. TF-KD benefits from temporal dependency through teacher forcing training, while SS-KD benefits from adequate data exposure through scheduled sampling strategy. MT-KD benefits from the both.

We observe in Table II that the performance gain of MT-KD over TF-KD/SS-KD is the most prominent, that is attributed to the multi-teacher strategy. Specifically, we can find that MT-KD achieves the lowest MCD and RMSE scores of 8.13 dB and 19.98 Hz for English that are lower than those of TF-KD and SS-KD. We observe the same trend for Chinese. To sum, 2-teacher knowledge distillation is more much effective than 1-teacher counterpart in addressing the exposure bias problem.

In summary, TF and SS training techniques each has its own advantages. The 2-teacher transfer learning benefits from the best of both, that effectively mitigates the exposure bias problem.

We note that MCD and RMSE measure the distortion of acoustic features. They do not reflect the perceptual quality of speech [66]. We next report the listening tests, in terms of mean opinion score (MOS) [67] and Best Worst Scaling (BWS) [68], [69] evaluations as the indicators of overall speech quality.

E. Naturalness evaluation

1) Subjective evaluation (MOS): We conduct the first listening experiment by reporting the MOS scores [67] across all systems, and summarize in the fifth and sixth columns of Table II.

Each speech sample is rated on a scale of 1 to 5 with an interval of 0.5. "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. We recruit 30 English and 30 Chinese listeners. Each listener listens to 200 speech samples of his/her native language. The listeners are instructed to pay attention to the naturalness of speech. We adopt both Griffin-Lim algorithm [37] and Parallel-WaveGAN vocoder [63] for speech waveform generation.

1.1) Experiment results

It is observed that data augmentation, adversarial learning, and knowledge distillation are all effective in addressing the exposure bias problem, outperforming the TF baseline. While TF-KD and SS-KD obtain comparable results, both of them consistently outperform TF-GAN, SS-GAN, and SS systems. Furthermore, MT-KD clearly stands out in both English and Chinese experiments, benefiting from multi-teacher and knowledge distillation strategy. Comparing MT-KD with others, we observe that its performance gain is attributed mainly to the 2-teacher strategy.

With Griffin-Lim waveform generator, MT-KD achieves a MOS of 3.82 for English, that is significantly higher than others; and 3.85 for Chinese. With Parallel-WaveGAN neural vocoder, MT-KD also outperforms the reference systems by a large margin, which is consistent with previous observations. It achieves a MOS of 4.09 for English and 4.12 for Chinese, that is very close to those of natural speech reference.

The MOS listening tests confirm that MT-KD outperforms other competing systems in terms of naturalness.

1.2) MT-KD training

Among the hyper-parameters that affect the MT-KD system performance, we would like to discuss the scaling factor λ in Eq. 17, that balances the contributions between $Loss_{d1}$ and $Loss_{d2}$. We perform a MOS listening experiment for 4 systems with $\lambda \in \{0.2, 0.4, 0.6, 0.8\}$. We follow the same experimental protocol as that for the MOS listening tests in Section IV-E1. Parallel-WaveGAN vocoder is used to synthesize speech.

Figure 5 summarizes the performance of MT-KD with different λ values. We can see that the performance peaks when λ = 0.4. We note that a higher value of λ does not lead to better performance. By adjusting λ, we can balance the contributions from the two teachers. Overall, we empirically find that λ = 0.4 is a proper choice for our multi-teacher knowledge distillation.

2) Subjective evaluation (BWS): We conduct the second listening experiment through BWS [68], [69], which is an effective method to provide a ranking of a long list of listening samples [70]. In so doing, we randomly select 80 utterances from the in-domain set for Chinese and English respectively. We also recruit 15 English and 15 Chinese listeners. For each utterance, seven speech samples produced by these seven systems form a group. A listener picks the best and worst samples in terms of naturalness for each group. In other words, each listener listens to all 80 groups, 560 utterances in total, of his/her native language. All speech samples are generated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2022.3171974, IEEE/ACM Transactions on Audio, Speech, and Language Processing

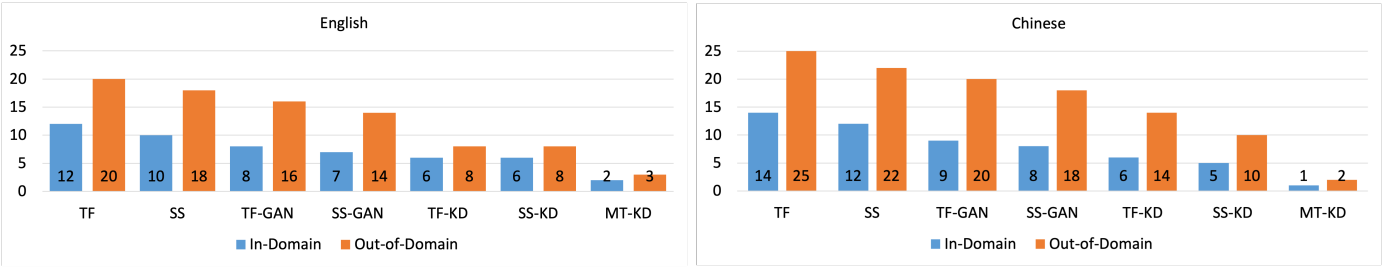

Fig. 6: The results of robustness evaluation in terms of Word Error Rate (WER %) in a listening experiment for both in-domain and out-of-domain utterances.

by Parallel-WaveGAN vocoder. We report the results in the last two columns of Table II.

We observe from Table II that TF is selected for 66% of time as the worst model for English, and 80% for Chinese. No listener indicates the TF as the best system. The results suggest that both knowledge transfer and data augmentation systems clearly outperform TF baseline. In general, the listeners favor knowledge transfer systems to data augmentation systems. Within knowledge transfer systems, the listeners prefer knowledge distillation systems over adversarial learning systems.

We also observe that MT-KD consistently outperforms all other systems, with 80% best votes for English and 86% best votes for Chinese. The results suggest that the listeners have a clear preference towards MT-KD system. The BWS gain by MT-KD over other systems is mainly attributed to the 2-teacher strategy. This is a strong indication that MT-KD fuses the knowledge of TF and SS to effectively enhance the decoding capability of SS for more natural speech synthesis.

F. Robustness evaluation

In Tacotron-based neural TTS, the exposure bias problem leads to unexpected errors during autoregressive inference [62], [71]. We further conduct a listening experiment for robustness evaluation and report the Word Error Rate (WER %), which reflects the robustness of speech [61]. WER is defined as the total number of errors, i.e., repetition, mispronunciation, and deletion, over the total number of words in a listening experiment. We follow the experiment protocol in [61], [62].

We recruit 15 English and 15 Chinese listeners to identify the errors [61], [62] in the synthesized utterances. We select 80 utterances from the in-domain set and 80 utterances from the out-of-domain set, for Chinese and English respectively. In other words, each listener listens to 160 utterances. In this experiment, we generate speech waveform with Parallel-WaveGAN vocoder. The word error rates (WER %) for seven systems are reported in Figure 6.

As shown in Figure 6, generally all systems perform better on in-domain test data than on out-of-domain test data. All knowledge transfer and data augmentation systems outperform the TF baseline. Furthermore, just like in the naturalness experiments, knowledge distillation systems outperform adversarial learning systems consistently.

Among the knowledge distillation systems, while TF-KD and SS-KD provide similar results, MT-KD clearly outperforms TF-KD and SS-KD, that highlights the advantage of 2-teacher strategy. For English, MT-KD achieves WER of 2% for in-domain set and 3% for out-of-domain set. For Chinese, MT-KD shows consistent performance and achieves WER of 1% for in-domain set and 2% for out-of-domain set.

The WER reported in this listening experiment is an indicator of perceptual quality of speech as perceived by human listeners. We are encouraged by the fact that robustness evaluation results in Figure 6 are highly correlated with the objective evaluation results in Table II.

G. Expressiveness evaluation

One of the typical symptoms of the exposure bias problem [33] in Tacotron2 TTS system is inappropriate prosody phrase breaking. Prosody phrase breaks separate a long utterance into a sequence of breath groups [72], that are semantically and syntactically appropriate in natural speech.

We use Break Error Rate (BER %) as the proxy of expressiveness of utterances. BER is defined as the ratio between the number of utterances with errors over the total number of utterances in the listening experiment. On the same evaluation data in Section IV-F, we report the expressiveness performance. The same 15 English and 15 Chinese listeners in Section IV-F are invited to identify incorrect phase breaks, i.e., errors, in the synthesized utterances. Each listener also accesses 160 utterances for their own language, including 80 utterances from the in-domain set and 80 utterances from the out-of-domain set. The Parallel-WaveGAN is used to synthesize the speech. The results are reported in Table III.

1) Experiment results

We observe that all knowledge transfer and data augmentation systems outperform the TF baseline. The scheduled sampling systems achieve lower BER than their teacher forcing counterparts, e.g., SS vs TF, SS-GAN vs TF-GAN, SS-KD vs TF-KD, in both in-domain and out-of-domain utterances, which suggests that scheduled sampling is effective in addressing the exposure bias problem. The knowledge distillation systems consistently show better prosodic phrase breaking results than their adversarial learning counterparts. Furthermore, MT-KD clearly outperforms all baselines for both in-domain and out-of-domain utterances. In particular, in the out-of-domain test, MT-KD achieves BER of 8% for English and 5% for Chinese, that is encouraging; in the in-domain test, our MT-KD achieves BER of 2% for English, and doesn't see any phrase breaking error for Chinese. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2022.3171974, IEEE/ACM Transactions on Audio, Speech, and Language Processing

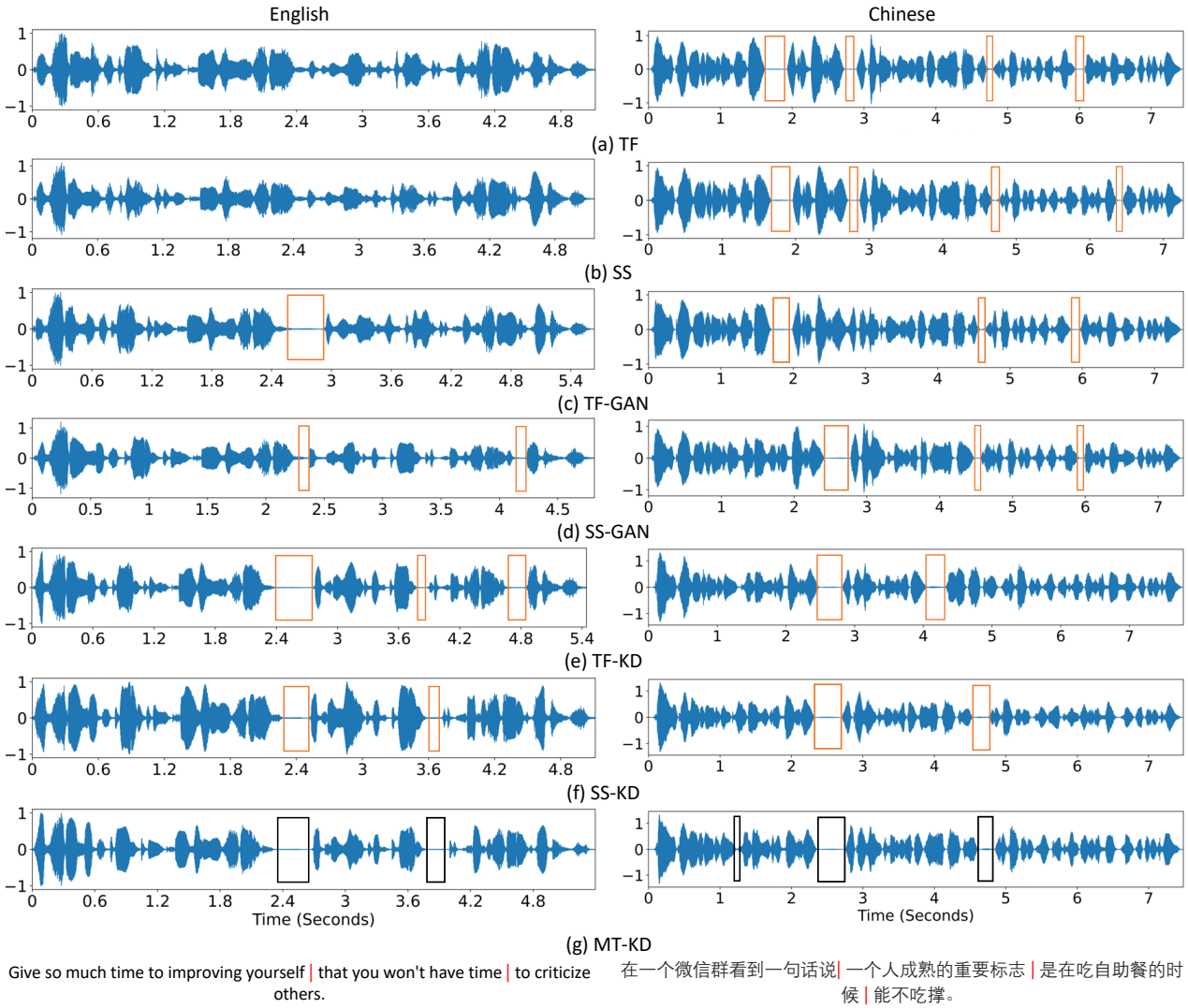

Fig. 7: Visualization of synthesized out-of-domain utterances for an English and a Chinese scripts. The orange boxes denote the prosodic phase breaks in the utterances. The black boxes for "(g) MT-KD" denote the predicted natural breaks in the utterances that coincide with the appropriate phrase breaks. The red bars in the scripts suggest the appropriate phrase breaks which are consistent with human intuition as can be verified at our demo site.

TABLE IV: The naturalness evaluation results in terms of Mean Opinion Score (MOS), the robustness evaluation results in terms of Word Error Rate (WER %), and the expressiveness evaluation results in terms of sentence-level Break Error Rate (BER %) for a listening experiment based on the GST-Tacotron network architecture.

System	Language	MOS		WER (%)	BER (%)
		GL	PW		
GST-TF	English	3.75 ± 0.04	3.94 ± 0.02	6	11
	Chinese	3.75 ± 0.03	3.96 ± 0.02	7	9
GST-SS	English	3.78 ± 0.04	3.95 ± 0.05	5	9
	Chinese	3.79 ± 0.04	3.99 ± 0.02	4	8
GST-TF-KD	English	3.80 ± 0.01	3.97 ± 0.03	3	4
	Chinese	3.84 ± 0.03	4.02 ± 0.02	2	5
GST-SS-KD	English	3.83 ± 0.02	4.05 ± 0.04	2	4
	Chinese	3.85 ± 0.04	4.06 ± 0.01	2	3
GST-MT-KD	English	**3.85 ± 0.01**	**4.10 ± 0.02**	**0**	**1**
	Chinese	**3.87 ± 0.03**	**4.13 ± 0.03**	**1**	**0**
Ground Truth	English	4.40 ± 0.01		NA	NA
	Chinese	4.42 ± 0.03		NA	NA

results confirm the individual contributions by multi-teacher and knowledge distillation.

2) Case study

We now provide a case study to illustrate the prosodic phrase breaking behaviors. Figure 7 shows a waveform plot of all comparative systems. It is clear that MT-KD produces better prosodic phrase breaks than others. We have added orange boxes to indicate the natural prosodic phrase breaks. For example, in the English utterance, we have "*Give so much time to improving yourself | that you won't have time | to criticize others.*" with the bars representing the natural breaks. Examples with different prosodic phrase breaking behaviors across different systems are available at our demo site.

In sum, all the above experiments confirm that the MT-KD system effectively addresses the exposure bias problem in naturalness, robustness and expressiveness evaluations. The performance gain of MT-KD over data augmentation and adversarial learning systems is attributed to the knowledge distillation strategy, and the performance gain over other knowledge distillation systems is attributed to the 2-teacher strategy.

V. Evaluation on GST-Tacotron

The proposed (multi-teacher) knowledge distillation framework is generally applicable to other network architectures. We further validate it on GST-Tacotron [73], the state-of-the-art TTS framework. The GST-Tacotron model [73] was originally designed for style control and transfer. It also attempts to address the run-time stability during speech generation, such as deletion or repetition of words, incomplete utterance, and inappropriate prosody phrase break, arising from the exposure bias problem.

We implement five training schemes with the GST-Tacotron model in a comparative study, and refer 'Tacotron' in GST-Tacotron to the Tacotron2 TTS framework.

- GST-TF: *GST-Tacotron* TTS model [73] trained in the teacher forcing mode.
- GST-SS : *GST-Tacotron* TTS model [73] trained in the scheduled sampling mode [27] for data augmentation.
- GST-TF-KD : *GST-Tacotron* TTS model [73] with 1-teacher knowledge distillation, i.e., *Teacher-TF* only.
- GST-SS-KD: *GST-Tacotron* TTS model [73] with knowledge distillation of 1-teacher, i.e., *Teacher-SS* only.
- GST-MT-KD: *GST-Tacotron* TTS model [73] with the proposed multi-teacher knowledge distillation scheme.

We only compare all systems under the parallel style transfer scenario [73] to observe the naturalness, robustness and expressiveness, because the non-parallel style transfer scenario [73] introduces extra varying factors, that make it difficult to draw conclusions. In the parallel style transfer scenario, the ground truth speech data is required as the reference. As the out-of-domain set does not have ground truth speech, we only select 100 utterances from the in-domain set as the evaluation utterances for Chinese and English respectively.

1) Naturalness evaluation

We generate speech waveform with Griffin-Lim algorithm [37] and Parallel-WaveGAN vocoder [63] and recruit 15 English and 15 Chinese listeners to conduct the naturalness evaluation in terms of MOS score. The MOS scores for five systems and ground truth speech are summarized in the third and fourth columns of Table IV.

We can observe that all knowledge distillation and data augmentation (GST-SS) systems achieve higher MOS scores, that are closer to the ground truth speech, than the GST-TF baseline. Furthermore, we're happy to see that GST-MT-KD outperforms GST-TF-KD and GST-SS-KD consistently, which validates the effectiveness of the multi-knowledge distillation.

2) Robustness evaluation

As shown in the fifth column of Table IV, all knowledge distillation and data augmentation (GST-SS) systems outperform the GST-TF baseline. Furthermore, just like in the former experiments, multi-knowledge distillation system, GST-MT-KD outperforms single teacher systems, namely, GST-TF-KD and GST-SS-KD, consistently.

3) Expressiveness evaluation

The expressiveness results are reported in the last column of Table IV. We observe that all knowledge distillation and data augmentation systems outperform the GST-TF baseline. The knowledge distillation systems consistently show better prosodic phrase breaking than others. Furthermore, GST-MT-KD obtains the best performance. All results suggest that knowledge distillation is effective in addressing the exposure bias problem.

It is encouraging to observe that, while the GST-Tacotron network architecture has addressed the run-time stability problem in its own way, the proposed MT-KD framework further improves its naturalness, robustness and expressiveness.

VI. Conclusion

We have studied a novel decoding knowledge transfer strategy and a multi-teacher knowledge distillation (MT-KD) scheme to address the exposure bias problem in neural TTS. In the experiments, we confirm our intuition that knowledge distillation in general is more effective than data augmentation and adversarial learning, and 2-teacher knowledge distillation outperforms 1-teacher counterpart by a large margin. MT-KD framework outperforms all reference systems in terms of naturalness, robustness and expressiveness. Further experiments show that our MT-KD method is also effective on the GST-Tacotron network architecture. The MT-KD is focused on transferring knowledge from a pre-trained teacher to a student model to mitigate the exposure bias issue, where the pre-trained teacher is not used at run-time inference. As a future work, we will investigate how to take advantage of the pre-trained models in the decoding process for multi-pass decoding [74].

References

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779–4783.

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[4] H. Zen, A. W. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962–7966.

[5] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. ISCA Speech Synthesis Workshop*, 2016, pp. 202–207.

[6] R. Liu, F. Bao, G. Gao, and Y. Wang, "Mongolian text-to-speech system based on deep neural network," in *Proc. National Conference on Man-Machine Speech Communication*, 2017, pp. 99–108.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Proc. ISCA Speech Synthesis Workshop*, 2016, pp. 125–130.

[8] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. International Conference on Machine Learning*, vol. 80, 2018, pp. 4700–4709.

[9] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "Wavetts: Tacotron-based TTS with joint time-frequency domain loss," in *Proc. The Speaker and Language Recognition Workshop*, 2020, pp. 245–251.

[10] Y. Zhou, X. Tian, and H. Li, "Language agnostic speaker embedding for cross-lingual personalized speech generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3427–3439, 2021.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. International Conference on Learning Representations ICLR*, 2015, pp. 1–10.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] M. I. Jordan, "Serial order: A parallel distributed processing approach," in *Advances in psychology*, 1997, vol. 121, pp. 471–495.

[14] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[15] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 712–718.

[16] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and waveglow or single gaussian wavernn vocoders," in *Proc. INTERSPEECH*, 2019, pp. 1308–1312.

[17] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial investigation of an encoder-decoder end-to-end tts framework using marginalization of monotonic hard latent alignments," in *Proc. ISCA Speech Synthesis Workshop*, 2019, pp. 1–6.

[18] J. Fong, P. O. Gallegos, Z. Hodari, and S. King, "Investigating the robustness of sequence-to-sequence text-to-speech models to imperfectly-transcribed training data." in *Proc. INTERSPEECH*, 2019, pp. 1546–1550.

[19] Y. Yasuda, X. Wang, and J. Yamagishi, "Effect of choice of probability distribution, randomness, and search methods for alignment modeling in sequence-to-sequence text-to-speech synthesis using hard alignment," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6724–6728.

[20] C. Valentini-Botinhao and S. King, "Detection and Analysis of Attention Errors in Sequence-to-Sequence Text-to-Speech," in *Proc. INTERSPEECH*, 2021, pp. 2746–2750.

[21] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," in *Proc. INTERSPEECH*, 2019, pp. 1293–1297.

[22] R. Liu, F. Bao, G. Gao, H. Zhang, and Y. Wang, "Improving mongolian phrase break prediction by using syllable and morphological embeddings with bilstm model," in *Proc. INTERSPEECH*, 2018, pp. 57–61.

[23] Y. Zheng, X. Wang, L. He, S. Pan, F. K. Soong, Z. Wen, and J. Tao, "Forward-backward decoding for regularizing end-to-end TTS," in *Proc. INTERSPEECH*, 2019, pp. 1283–1287.

[24] R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li, "Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 274–285, 2021.

[25] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. International Conference on Learning Representations ICLR*, 2016, pp. 1–10.

[26] F. Schmidt, "Generalization in generation: A closer look at exposure bias," in *Proc. Workshop on Neural Generation and Translation*, 2019, pp. 157–167.

[27] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[28] F. Huszár, "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?" *arXiv preprint arXiv:1511.05101*, 2015.

[29] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[30] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, "Professor forcing: A new algorithm for training recurrent networks," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 4601–4609.

[31] A. Venkatraman, M. Hebert, and J. A. Bagnell, "Improving multi-step prediction of learned time series models," in *Proc. AAAI Conference on Artificial Intelligence*, 2015, pp. 3024–3030.

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[33] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based TTS," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6274–6278.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[35] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Transactions on Neural Networks Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[37] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[38] K. E. Ak, N. Xu, Z. Lin, and Y. Wang, "Incorporating reinforced adversarial learning in autoregressive image generation," in *Proc. European Conference Computer Vision (ECCV)*, vol. 12366, 2020, pp. 18–34.

[39] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1179–1195.

[40] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, "Improving image captioning with conditional generative adversarial nets," in *Proc. AAAI Conference on Artificial Intelligence*, 2019, pp. 8142–8150.

[41] D. Liu, Z. Zha, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for sequence-level image captioning," in *Proc. ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 1416–1424.

[42] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 4334–4343.

[43] T. Wang, X. Liu, J. Tao, J. Yi, R. Fu, and Z. Wen, "Non-autoregressive end-to-end TTS with coarse-to-fine decoding," in *Proc. INTERSPEECH*, 2020, pp. 3984–3988.

[44] H. Guo, F. K. Soong, L. He, and L. Xie, "A new gan-based end-to-end TTS training algorithm," in *Proc. INTERSPEECH*, 2019, pp. 1288–1292.

[45] T. Miyato, A. M. Dai, and I. J. Goodfellow, "Adversarial training methods for semi-supervised text classification," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
[46] L. Wu, Y. Xia, F. Tian, L. Zhao, T. Qin, J. Lai, and T. Liu, "Adversarial neural machine translation," in *Proc. Asian Conference on Machine Learning, (ACML)*, vol. 95, 2018, pp. 534–549.
[47] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5024–5028.
[48] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.
[49] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1317–1327.
[50] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born-again neural networks," in *Proc. International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 1602–1611.
[51] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2018, pp. 4320–4328.
[52] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2021.
[53] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," in *Proc. INTERSPEECH*, 2019, pp. 1128–1132.
[54] J. W. Yoon, H. Lee, H. Y. Kim, W. I. Cho, and N. S. Kim, "Tutornet: Towards flexible knowledge distillation for end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1626–1638, 2021.
[55] Y. Liu, W. Zhang, and J. Wang, "Adaptive multi-teacher multi-level knowledge distillation," *Neurocomputing*, vol. 415, pp. 106–113, 2020.
[56] K. Wang, Y. Liu, Q. Ma, and Q. Z. Sheng, "Mulde: Multi-teacher knowledge distillation for low-dimensional knowledge graph embeddings," in *Proc. Web Conference*, 2021, pp. 1716–1726.
[57] M. Wu, C. Chiu, and K. Wu, "Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2202–2206.
[58] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *proc. Advances in Neural Information Processing Systems*, 2016, pp. 2226–2234.
[59] L. Prechelt, "Early stopping-but when?" in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science, 1996, vol. 1524, pp. 55–69.
[60] R. Liu, J. Li, F. Bao, and G. Gao, "The imu speech synthesis entry for blizzard challenge 2019," *Proc. Blizzard Challenge*, vol. 2019, 2019.
[61] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Advances in Neural Information Processing Systems*, 2019, pp. 3165–3174.
[62] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *Proc. International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 7586–7598.
[63] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6199–6203.
[64] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.
[65] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.
[66] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.
[67] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.
[68] J. A. Lee, G. Soutar, and J. Louviere, "The best–worst scaling approach: An alternative to schwartz's values survey," *Journal of personality assessment*, vol. 90, no. 4, pp. 335–347, 2008.
[69] J. J. Louviere, T. N. Flynn, and A. A. J. Marley, *Best-worst scaling: Theory, methods and applications*, 2015.
[70] T. N. Flynn and A. A. Marley, "Best-worst scaling: theory and methods," in *Handbook of choice modelling*, 2014.
[71] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *proc. International Conference on Learning Representations (ICLR)*, 2018.
[72] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "Modeling prosodic phrasing with multi-task learning in tacotron-based tts," *IEEE Signal Processing Letters*, vol. 27, pp. 1470–1474, 2020.
[73] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 5167–5176.
[74] S. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proc. AAAI Conference on Artificial Intelligence*, 2020, pp. 5191–5198.

Rui Liu received his B.S. degree from the Department of Software at Taiyuan university of technology, Taiyuan, China, in 2014, and the PhD degree in Computer Science and Technology from Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Inner Mongolia University, Hohhot, China, in 2020. He is also an exchange PhD candidate at the Department of Electrical & Computer Engineering of National University of Singapore (NUS), funded by China Scholarship Council (CSC). He is currently working as a joint Postdoctoral Research Fellow at National University of Singapore (NUS) and Singapore University of Technology and Design (SUTD). His research interests include prosody and acoustic modeling for speech synthesis, machine learning and natural language processing.

Berrak Sisman received her PhD degree in Electrical and Computer Engineering from National University of Singapore in 2020, fully funded by A*STAR Graduate Academy under Singapore International Graduate Award (SINGA). She is currently working as an Assistant Professor at Singapore University of Technology and Design (SUTD). She is also an Affiliated Researcher at the National University of Singapore (NUS). Prior to joining SUTD, she was a Postdoctoral Research Fellow at the National University of Singapore, and a Visiting Researcher at Columbia University, New York, United States. She was also an exchange PhD student at the University of Edinburgh and a visiting scholar at The Centre for Speech Technology Research (CSTR), University of Edinburgh in 2019. She was attached to RIKEN Advanced Intelligence Project, Japan in 2018. Her research is focused on machine learning, signal processing, speech synthesis and voice conversion. She has served as the General Coordinator of the Student Advisory Committee (SAC) of International Speech Communication Association (ISCA).

Guanglai Gao received the B.S. degree from Inner Mongolia University, Hohhot, China, in 1985, and the M.S. degree from the National University of Defense Technology, Changsha, China, in 1988. He was a Visiting Researcher at University of Montreal, Canada. Currently, he is a Professor with the Department of Computer Science, Inner Mongolia University. His research interests include artificial intelligence and pattern recognition.

Haizhou Li Haizhou Li (M'91-SM'01-F'14) received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at the School of Data Science, the Chinese University of Hong Kong, Shenzhen, China, and the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining NUS, he taught in the University of Hong Kong (1988-1990) and South China University of Technology (1990-1994). He was a Visiting Professor at CRIN in France (1994-1995), Research Manager at the Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), Vice President in InfoTalk Corp. Ltd. (2001-2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003-2016). Dr Li served as the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2018), a Member of the Editorial Board of Computer Speech and Language (2012-2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013-2015), the President of the International Speech Communication Association (2015-2017), the President of Asia Pacific Signal and Information Processing Association (2015-2016), and the President of Asian Federation of Natural Language Processing (2017-2018). He was the General Chair of ACL 2012, INTERSPEECH 2014 and ASRU 2019. Dr Li is a Fellow of the IEEE and the ISCA. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, U Bremen Excellence Chair Professor in 2019, and Fellow of Academy of Engineering Singapore in 2022.