Multistage Deep Transfer Learning for EmIoT-Enabled Human–Computer Interaction

Rui Liu, *Member, IEEE*, Qi Liu, *Member, IEEE*, Hongxu Zhu, *Member, IEEE*, and Hui Cao, *Member, IEEE*

***Abstract*—Emotional Internet of Things (EmIoT), which provides Internet of Things (IoT) devices cognitive and socialization capabilities, has been regarded as a future direction to improve users' experiences. With the development of intelligent techniques, the requirement of EmIoT is not only sensing the users' emotional states but also providing emotional feedbacks. Human–computer interaction has been studied to achieve speech interaction with IoT devices. The recent advances in neural text-to-speech (TTS) have made "human parity" synthesized speech possible for IoT-enabled human–computer interaction. Furthermore, emotion control can be achieved by using the emotional codes in a unified model, referred to as emotional TTS (or ETTS for short). Such ETTS models have achieved promising emotional expressiveness using large-scale emotion-annotated English data set; however, they are not practical in IoT environments with other mainstream languages, especially for Chinese. In fact, the limited available large-scale emotion-annotated data set is challenging the development of Chinese ETTS. To address that we propose a multistage deep transfer learning scheme to design a high-quality Chinese ETTS system under a small-scale training corpus to achieve EmIoT in Mandarin environments. In this scheme, the pretrained knowledge from the former stages corresponding to a large-scale neutral English and a medium-scale emotional English corpora is transferred to a Mandarin ETTS model. Thereby, the trained model can achieve high-quality emotional speech with limited available emotional corpus, which is able to serve various EmIoT-oriented applications. The experiments have been conducted to demonstrate the effectiveness and superiority of the proposed model as compared to other counterparts in terms of naturalness and emotional expressiveness. We refer readers to visit our demo Webpage[1] and enjoy the synthesized speech samples.**

***Index Terms*—Emotional expressiveness, emotional Internet of Things (EmIoT), human–computer interaction (HCI), transfer learning.**

Manuscript received 3 September 2021; revised 16 December 2021; accepted 19 January 2022. Date of publication 3 February 2022; date of current version 8 August 2022. This work was supported by the High-Level Talents Introduction Project of Steed Program of Inner Mongolia University (presided over by Rui Liu). *(Corresponding author: Qi Liu.)*

Rui Liu is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: liurui_imu@163.com).

Qi Liu is with the School of Future Technology, Guangzhou International Campus, South China University of Technology, Guangzhou 511442, Guangdong, China (e-mail: qiliu47-c@my.cityu.edu.hk).

Hongxu Zhu is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore (e-mail: hxzhu@nus.edu.sg).

Hui Cao is with the School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China (e-mail: tsaohui@163.com).

Digital Object Identifier 10.1109/JIOT.2022.3148766

[1] *Synthesized Speech Samples:* https://ttslr.github.io/IOT/.

I. Introduction

During the past decades, the information exchange network, with the Internet of Things (IoT) as the backbone, has generated explosive smart applications [1], [2], such as smart home [3], smart healthcare [4], etc. With the development of intelligent technologies, the target toward IoT has far beyond collecting data. For instance, the current expectation on smart home is not only collecting the environmental data but also automatically operating home devices (e.g., air conditioner and lights) to adaptively generate a comfortable room [5]. Ambitiously, it is also essential to create a connected life that achieves the interaction between humans and IoT devices [6]. Emotional IoT (EmIoT) is an emerging framework that identifies a possible direction for IoT under this trend.

The main objective of EmIoT is giving emotional intelligence to the IoT [7]. Specifically, IoT systems should be able to feel, understand the emotion of the user, and try to generate a level of affection based on the interaction between the IoT world and the user. Although such an interaction is bidirectional, most of the existing works [8], [9] focus on understanding users' emotions, whereas the emotional feedback from IoT to the user is rarely covered.

As one main modality that humans utilize to communicate with each other, speech is one of the most convenient means to generate interactions [10]. Thus, voice user interface (VUI) [11] has obtained significant superiority over the confusing and difficult graphical user interface (GUI) [12]. Consequently, recent advances in bidirectional human–computer interaction (HCI) have been studied to achieve speech interaction using IoT [9], [10], [13]. The traditional bidirectional speech communication for IoT consists of speech recognition, dialogue generation, and text-to-speech (TTS) or speech synthesis modules [14]. The TTS technology is becoming an integral part of interacting with IoT.

As a sequence-to-sequence mapping problem, TTS synthesis aims to render a naturally sounding speech waveform to be synthesized from a text [15], [16], that is, from a sequence of discrete symbols (viz., texts) to a real valued time series (viz., waveforms). Nowadays, TTS systems are mainly divided into threefold: 1) concatenative [17], [18]; 2) statistical parametric [16]; and 3) end-to-end speech synthesis [19]–[26]. However, the former two TTS systems require complex pipeline, including text preprocessing [16], duration model [18], acoustic model [18], waveform generation [16], etc., to achieve TTS synthesis, rendering the limited application in IoT. With the advent of deep learning,

end-to-end generative TTS models simplify the synthesis pipeline with a single neural network, and the representatives include Tacotron [19], Tacotron2 [20], and their variants [21]–[26]. The rationales of them are to integrate the conventional TTS pipeline into a unified encoder–decoder network, and then to learn the mapping directly from the <text, wav> pair [19], [20]. Furthermore, together with a neural vocoder [27]–[30], natural-sounding human-like speech can be generated.

Notwithstanding the advances in TTS, the emotional expressiveness of the synthesized speech remains to be improved [21], [23] to achieve an emotional feedback in EmIoT. Many works have been proposed to apply the global style tokens (GSTs) to control the expressive effect, such as [21], [23], and [24], which are referred to as the GST-Tacotron paradigm. Nevertheless, how to generate speech with expected emotions is becoming an important research topic in TTS. To that end, some works utilize emotional states as control vectors to effectively generate emotional speech [31]–[35]. It is worth noting that deep learning-based emotional TTS (ETTS) on English data set has achieved successfully due to the available large amount of emotional English TTS corpus [36]. Unfortunately, owing to the cost expensive for recording, collecting, and processing large-scale, high-quality emotional Mandarin TTS corpus, it necessitates the end-to-end neural TTS framework to still work well with limited public emotional speech corpus [37], [38]. Given the huge proportion of Mandarin environment in the IoT market [39], therefore, to design a high-quality emotional Mandarin TTS scheme becomes necessary.

In fact, for the Mandarin ETTS task, the lack of large-scale emotion-annotated training data becomes more rigorous, leading to unsatisfactory results [40]. Recently, the transfer learning strategy is employed to overcome the limited data problem for Mandarin ETTS [40], [41]. Furthermore, Wu *et al.* [40] fine-tuned a Mandarin ETTS model based on a pretrained TTS model with limited data. However, the fine-tuning method, that is, a simple model adaptation-based transfer learning method, cannot adapt general purposes for knowledge representation to be task aware on a downstream task, which leads to the unsmooth transferring issue and therefore, damages the performance [42]. The question is how to design a deep transfer learning method to take advantage of the limited data to achieve smooth transferring for high-quality Mandarin ETTS performance.

Note that cross-lingual transfer learning has shown its promising learning capability in the target domain from out-of-domain data, and further boosting performance [25], [43]. Many studies point to the fact that emotion is language universal to some extent [44]–[46]. Therefore, the emotional clues in English pronunciation are utilized to provide useful knowledge for emotional Mandarin TTS.

Inspired by this, to address the unsmooth transferring issue of the simple model adaptation-based transfer learning method, we design a multistage cross-lingual knowledge transfer strategy to ensure that the emotional knowledge learned from the former stage can be transferred to the latter stage smoothly between different stages. In this article, we formulate a novel cross-lingual transfer learning strategy to make the most of limited Mandarin data for emotional Mandarin TTS. We consider the cross-lingual transfer learning problem for emotional Mandarin speech generation, where a source language (English) task with plentiful ETTS data is utilized to improve the performance of an ETTS model on a target language (Mandarin) task with limited available ETTS corpus. This work makes notable contributions in four areas that are as follows.

1) A novel multistage deep transfer learning scheme is proposed to address the unsmooth transferring issue of simple model adaptation-based transfer learning for Mandarin ETTS.
2) A graceful deep emotion knowledge transfer process is proposed by using three-stage cross-lingual transfer learning from English to Mandarin languages. Note that to the best of our knowledge, this is not investigated in any prior works for emotional Mandarin TTS task.
3) The proposed method is superior to the state-of-the-art ETTS baselines in both spectrum and emotion modeling, with the advantage of projecting high-quality emotional Mandarin speech under the limited resource scenario, especially in EmIoT.
4) A novel high-quality emotional Mandarin TTS system is developed with limited emotional corpus to support the emotional feedback requirement in EmIoT.

The remainder of this article is organized as follows. In Section II, we discuss the background to motivate our research. In Section III, we propose a novel training strategy for the emotional Mandarin TTS system with multistage transfer learning. The experimental results are reported in Section IV. Finally, we conclude in Section V.

II. Background

We first brief the EmIoT and review the bidirectional HCI for IoT, followed by the description of the ETTS model and transfer learning.

A. EmIoT and Bidirectional Speech Communication for IoT

Emotion plays an important role in human's daily health. Emotion detection, namely, the investigation on making a computer to feel users' emotions, has started since for a long time. With the development of IoT, such an emotion sensing function has also expected in various smart applications, such as smart home and smart healthcare. As an emerging framework, EmIoT raises a higher requirement on detecting users' emotion conditions and providing feedback for HCI to realize the VUI, which makes us directly speak to a device. Bidirectional speech communication allows the IoT devices to listen the users' commands and to give feedback with the human-like speech accordingly.

As shown in Fig. 1, the traditional bidirectional speech communication is built on a cascade framework [23], which consists of speech recognition, dialogue generation, and speech synthesis modules. The speech recognition module takes the speech command of end user as input and generates the text representation. The dialogue generation module seeks to understand the intention of the speech command and to response properly. The speech synthesis module transforms

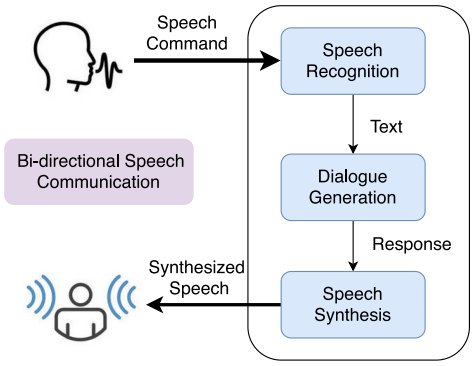

Fig. 1. Diagram of bidirectional speech communication for IoT.

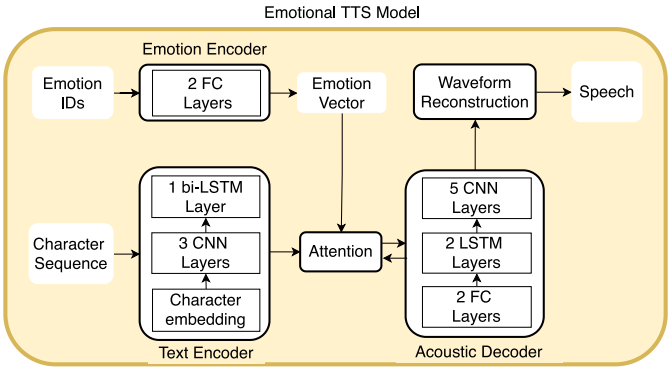

Fig. 2. Block diagram of the basic ETTS model. The model takes character sequence and the given emotion labels as input and outputs the corresponding mel-spectrum, which is then fed to the waveform reconstruction module (Griffin–Lim reconstruction algorithm [54] or WaveNet-like neural vocoder [27], [28], [55]) to synthesize emotional speech with specific emotions.

the generated response to a human-like speech. Generally, the tone of voice can directly reflect the speakers' emotions. Thus, several studies [47]–[49] have adopted the speech signal on emotion detection or recognition in IoT applications. However, these works only solve one-directional requirement in EmIoT, and the emotional feedback is still not addressed yet. Note that the speech synthesis module works as the last step in the workflow. Therefore, the overall performance of synthesized speech directly determines the qualities of the interaction feedback and users' experiences. End-to-end TTS systems [19], [20], [23], [25] greatly improve the voice quality of the synthesized speech. However, how to generate speech with expected emotions, especially in the Mandarin environment, is also an important topic to be addressed in TTS.

B. Emotional Text-to-Speech Generation

ETTS seeks to synthesize human-like natural-sounding voice for a given input text with desired emotional expression. Such an emotional feedback can achieve bidirectional requirement in EmIoT.

The early studies of ETTS are based on hidden Markov models [50]–[52]. For example, we can synthesize speech with a desired emotion through model interpolation [51] or by incorporating unsupervised expression cluster during training [52]. Recently, deep learning contributes to the ETTS [31], [53], where emotion codes can be used as control vectors to change TTS output. The Tacotron-based ETTS system [32], [40] is one of the successful implementations, as illustrated in Fig. 2. It consists of a text encoder, emotion encoder, attention-based decoder, and waveform reconstruction module.

The text encoder takes the character sequence of given input text as input and outputs the high-level feature representation [19], [20]. It is composed of three convolution neural network (CNN) layers, and a bidirectional LSTM (bi-LSTM) module. Moreover, the emotion encoder converts the given emotion ID (e.g., happy, sad, angry, etc.) to a fix-length emotion vector, resulting in the clear expressiveness of emotion attitude via a linear projection with two fully connected (FC) layers. Both high-level feature representation and emotion vectors are fed to the attention-based decoder to predict the mel-spectrum features frame by frame. The decoder consists of a 2-layer FC, two LSTM layers, and five CNN layers. The location-sensitive attention [56] is applied to learn the linguistic-acoustic alignment, and two popular algorithms, such as Griffin and Lim [54], and WaveNet-based neural vocoder [27], can be used to reconstruct the speech waveform from mel-spectrum features.

C. Transfer Learning

Transfer learning focuses on storing knowledge gain while solving one problem and applying it to a different but related problem [57]. A common strategy of transfer learning is pretraining a model on one data set, which consists of a large number of labeled samples, such as ImageNet [58], and then transferring the pretrained model to the target data set for fine-tuning [59]. It is greatly useful and important when encountering very limited training samples [60]. Other transfer learning strategies include automatic speech recognition (ASR) [57], [61], natural machine translation (NMT) [62], and so on. Furthermore, cross-lingual transfer learning is viable to build robust models for a low-resource target language by leveraging labeled data from other (source) languages [63]–[66]. For example, Lee and Lee [66] proposed several cross-lingual transfer learning approaches for question–answer (QA) task, where experiments are conducted using English data set as source language while to achieve new state of the art on a small Chinese QA data set.

Motived by the effectiveness of the cross-lingual transfer learning, in this article, we design a multistage cross-lingual knowledge transfer strategy to ensure that the emotional knowledge learned from the former stage can be transferred to the latter stage smoothly between different stages.

The notations used in our proposed method are summarized in Table I. We will describe our method in detail in the next section.

III. Multistage Deep Transfer Learning for Mandarin Emotional TTS

We propose a multistage deep transfer learning strategy for Mandarin ETTS in Fig. 3, which employs a three-stage training scheme to solve the data limitation issue for Mandarin ETTS.

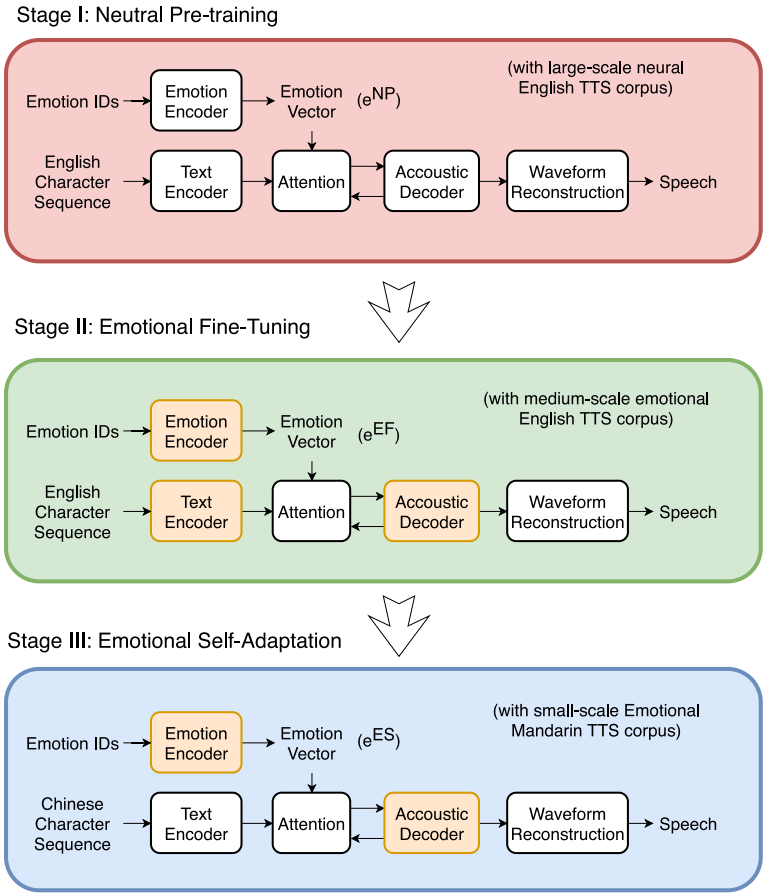

Fig. 3. Block diagram of the proposed multistage deep transfer learning system for emotional Mandarin TTS model. Stage-I is the NP stage, which trains the basic ETTS model using the large-scale English TTS corpus with neutral emotion. In Stage-II, the EF fine-tunes the pretrained ETTS model on the basis of the medium-scale English TTS corpus via three emotions (neural, happy, and sad). Meanwhile, the parameters of text encoder, emotion encoder, and acoustic decoder from Stage-I are transferred to that in Stage-II. The ES is in Stage-III, which is in charge of fine-tuning the ETTS model exploiting a small-scale emotional Mandarin TTS corpus with three emotions (neutral, happy, and sad). Note that the parameters from emotion encoder and acoustic decoder in Stage-II are mapped to Stage-III, respectively. Yellow boxes in Stages II and III mean that their parameters are initialized by the pretrained parameters from previous stage, while white boxes refer to that their parameters are initialized randomly.

TABLE I
DESCRIPTION OF NOTATIONS

Notation	Definition
D	The training set
x	The English character sequence
y	The target mel-spectrum sequence
h	The encoder hidden state
T	The length of English character sequence
$T^{'}$	The length of mel-spectrum sequence
l	The emotion label
e	The emotion vector
s	The decoder hidden state
c	The attention vector
α	The attention weight
w, W, V, U, F	The weight matrices
f, b	The learnable variable
Θ	The model parameters
$\bigtriangledown$	The back propagation process

As shown in Fig. 3, the proposed three-stage deep transfer learning scheme consists of *Stage I:* neutral pretraining (NP), *Stage II:* emotional fine-tuning (EF), and *Stage III:* emotional self-adaptation (ES) stages. In the first NP stage, a large-scale English TTS corpus is used to learn the initial end-to-end TTS model parameters as a prior information for the second EF step. Whereby, the ETTS model is initialized with the pretrained neutral TTS model parameters and then trained with a medium-scale emotional English TTS corpus. In the last ES step, the ETTS model is initialized by the fine-tuned ETTS model parameters from the former step and trained with a small-scale emotional Mandarin TTS corpus.

A. Stage I: Neutral Pretraining

In stage I, the text encoder aims to extract robust and efficient sequential hidden representation from input text. Specifically, the input English character sequence is a continuous real-valued vector, denoted as *character embedding*. Then, the embedding sequence is processed by three CNN layers to extract a longer term context for feeding a single bi-LSTM layer, thereby to generate the hidden representation of input sequence.

Given an input English character sequence, denoted as $x = (x_1, x_2, \ldots, x_T)$ and the corresponding target mel-spectrum features $y = (y_1, y_2, \ldots, y_{T'})$, the text encoder, termed as TEnc^{NP}, reads x and outputs a hidden state h_t per t step

$$h_t = \text{TEnc}^{\text{NP}}(h_{t-1}, x_t) \tag{1}$$

in which $t \in [1, T]$, T is the length of the input character sequence, and T' means the length of output acoustic features.

The emotion encoder contains two FC layers, and its function is to convert the input emotion labels (l) to a continuous, fixed-length, and two-dimensional (2-D) emotion vector e^{NP}. That is

$$\mathrm{e}^{\mathrm{NP}} = \mathrm{EEnc}^{\mathrm{NP}}(l). \tag{2}$$

The acoustic decoder consumes the hidden representation of input sequence with the help of a location-sensitive attention network [20]. Then, the resulting mel-spectrum frame is first passed through two FC layers. After that, the output of the FC module is concatenated with the previous context vector to feed a stack of two LSTM layers. Next, the LSTM output combines with the attention context vector to feed five CNN layers, leading to the underlying mel-spectrum output. For the attention-based decoder, viz., $\mathrm{Dec}^{\mathrm{NP}}$, the emotion vector e^{NP}, the previous hidden state s_{t-1}, and output y_{t-1} together with an attention vector c_t are taken as its input. Therefore, the mel-spectrum features y_t are obtained as

$$y_t = \mathrm{Dec}^{\mathrm{NP}}\left(y_{t-1}, s_{t-1}, e^{\mathrm{NP}}, c_t\right) \tag{3}$$

where the context vector c_t is calculated by

$$c_t = \sum_{i=1}^{T} \alpha_{t,i} h_i \tag{4}$$

in which $\alpha_{t,i}$ is the attention weight of h_i, and h_i represents the encoder output. The attention scores, a.k.a., values of the attention weights, help the network to focus on different parts of the input sequence. There are many choices for the implementation of the score function. In this article, we adopt the location sensitive attention mechanism [20] for the calculation of $\alpha_{t,i}$

$$\alpha_{t,i} = \mathrm{softmax}\left(e_{t,i}\right) \tag{5}$$

$$e_{t,i} = w^T \tanh\left(\mathbf{W}s_{t-1} + \mathbf{V}h_i + \mathbf{U}f_{t,i} + b\right) \tag{6}$$

$$f_t = \mathbf{F} * \alpha_{t-1} \tag{7}$$

where $*$ denotes 1-D convolution. W, V, U, and F are weight matrices and they are trained in conjunction with the entire TTS framework. α_{t-1} is the attention weight in the previous decoder step.

During the pretraining stage, the parameters of text encoder, emotion encoder, and acoustic decoder are updated simultaneously. It is assumed that the trained text encoder $\mathrm{TEnc}^{\mathrm{NP}}$ and decoder $\mathrm{Dec}^{\mathrm{NP}}$ create a constriction in the network that forces the information pertinent to emotion state into high-quality mel-spectrum features. The trained emotion encoder $\mathrm{EEnc}^{\mathrm{NP}}$ can clearly learn the emotion attitude expressed by the neutral speech. Note that the system on this stage just work on neutral TTS corpus to produce "neutral" emotion speech. The emotion vector e^{NP} generated by the emotion encoder $\mathrm{EEnc}^{\mathrm{NP}}$ is utilized to indicate the "neutral" emotion state. Hence, the emotion vector e^{NP} generated by the emotion encoder $\mathrm{EEnc}^{\mathrm{NP}}$ can express the neural emotion clues clearly. More importantly, for multistage transfer learning, the emotion encoder $\mathrm{EEnc}^{\mathrm{NP}}$ trained with neutral emotion speech corpus provides a warm and moderately starting point for the fine-tuning of various emotions (neutral, happy, and sad).

B. Stage II: Emotional Fine-Tuning

In the second EF stage, the ETTS model is initialized by the pretrained $\mathrm{TEnc}^{\mathrm{NP}}$, $\mathrm{EEnc}^{\mathrm{NP}}$, and $\mathrm{Dec}^{\mathrm{NP}}$, and then learned from a medium-scale emotional English TTS corpus.

As the same with the *Neural Pretraining*, the ETTS model shares the same network architecture, and the difference among them is the emotion encoder $\mathrm{EEnc}^{\mathrm{EF}}$. In this stage, we set the emotion labels to *neutral*, *happy*, and *sad* to match the medium-scale emotional English TTS corpus. The parameters of the emotion encoder $\mathrm{EEnc}^{\mathrm{EF}}$, text encoder $\mathrm{Enc}^{\mathrm{EF}}$, and the decoder $\mathrm{Dec}^{\mathrm{EF}}$ are initialized by the pretrained $\mathrm{EEnc}^{\mathrm{NP}}$, $\mathrm{Enc}^{\mathrm{NP}}$, and $\mathrm{Dec}^{\mathrm{NP}}$, respectively, and since then all parameters are updated in the process of fine-tuning. The motivation of this stage is to distinguish each emotional patterns corresponding to emotion labels. Meanwhile, the decoder $\mathrm{Dec}^{\mathrm{EF}}$ can generate both natural and emotional acoustic features.

C. Stage III: Emotional Self-Adaptation

On the basis of initialization from the former two stages, the ES stage contributes to training the emotional Mandarin TTS model ($\mathrm{TEnc}^{\mathrm{ES}}$, $\mathrm{EEnc}^{\mathrm{ES}}$, and $\mathrm{Dec}^{\mathrm{ES}}$) via a small-scale emotional Mandarin TTS corpus. Both systems in all stages own the same network architecture, however, the symbol list is revised to be able to process Chinese character, where the emotion labels are *neutral*, *happy*, and *sad* matching the Mandarin ETTS corpus. The details of the proposed multistage cross-lingual transfer learning algorithm are given in Algorithm 1.

With such three-stage processes, the emotional Mandarin TTS model is expected to learn the true probability distribution from nature English emotional clues, which is pretty informative for the Mandarin emotional speech projection due to the emotional language-universal property [44]–[46].

D. Model Deployment for IoT

At runtime, only the trained model of Stage III is involved, where the trained emotional Mandarin TTS model takes Chinese text as input to project emotional Mandarin speech. Fig. 4 shows the deployment details of the trained emotional Mandarin TTS model under the IoT environment.

The deployment workflow of the trained model consists of: 1) uplink phase; 2) processing phase; and 3) downlink phase. In the uplink phase, the end devices (e.g., smartphone, dialogue robot, car voice assistant, etc.) receive the speech command from users. The microphone inside the VUI of the end devices will collect the speech signal and forward the packet to a communication module. The communication module will send the speech signal to the central server through one or more IoT routers/gateways. Generally, given the short memory and consumption of the end devices, the trained emotional Mandarin TTS model and other related models (such as speech recognition, dialogue generation, etc.) are deployed in the central server with high computational power. In the processing phase, the user speech signal is preprocessed by speech recognition and understood by dialogue generation models, and the Mandarin ETTS mode is responsible for generating the emotional speech signal to provide emotional feedback to the user. In the downlink phase, the generated speech signal will be sent back to

Algorithm 1: Pseudocode of the Proposed Method

Input:
Training set:
$D = \{x, y, l\}$
x: character sequence
y: mel-spectrum feature sequence
l: emotion labels of emotional speech corpus
Output:
Θ: TTS model including TEnc, EEnc and Dec
Hyper-parameter:
N: epoch number
n: batch size
η: learning rate
Begin
> Stage-I: Neutral Pre-training
1: Initialize TTS model Θ^{NP}
2: **for** each iteration **do**
3: **for** each batch **do**
4: output $\hat{y}$:
$\hat{y} = \Theta^{NP}(x, l)$
5: update Θ^{NP} with MSE Loss :
$\Theta^{NP} \leftarrow \nabla_{\Theta^{NP}}(Loss(y, \hat{y}))$
6: **end for**
7: **end for**
8: return Θ^{NP} (namely, TEnc^{NP}, EEnc^{NP} and Dec^{NP})
> Stage-II: Emotional Fine-tuning
1: Initialize TTS model Θ^{EF}:
$\text{TEnc}^{\text{EF}} \leftarrow \text{TEnc}^{\text{NP}}$
$\text{EEnc}^{\text{EF}} \leftarrow \text{EEnc}^{\text{NP}}$
$\text{Dec}^{\text{EF}} \leftarrow \text{Dec}^{\text{NP}}$
2: **for** each iteration **do**
3: **for** each batch **do**
4: output $\hat{y}$:
$\hat{y} = \Theta^{EF}(x, l)$
5: update Θ^{EF} with MSE Loss :
$\Theta^{EF} \leftarrow \nabla_{\Theta^{EF}}(Loss(y, \hat{y}))$
6: **end for**
7: **end for**
8: return Θ^{EF} (viz., TEnc^{EF}, EEnc^{EF} and Dec^{EF})
> Stage-III: Emotional Self-Adaptation
1: Initialize TTS model Θ^{ES}:
$\text{EEnc}^{\text{ES}} \leftarrow \text{EEnc}^{\text{EF}}$
$\text{Dec}^{\text{ES}} \leftarrow \text{Dec}^{\text{EF}}$
2: **for** each iteration **do**
3: **for** each batch **do**
4: output $\hat{y}$:
$\hat{y} = \Theta^{ES}(x, l)$
5: update Θ^{ES} with MSE Loss :
$\Theta^{ES} \leftarrow \nabla_{\Theta^{ES}}(Loss(y, \hat{y}))$
6: **end for**
7: **end for**
8: return Θ^{ES} (i.e., TEnc^{ES}, EEnc^{ES} and Dec^{ES})
End

the end devices through the IoT routers/gateways. Finally, the speech signal will be played by the loudspeaker inside the VUI. Hence, the emotional feedback is achieved in EmIoT.

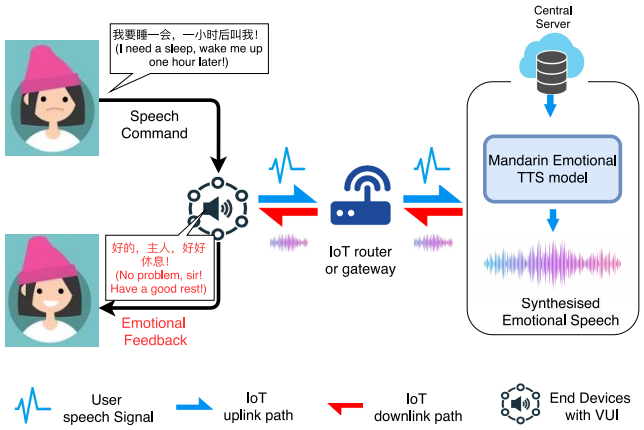

Fig. 4. Deployment of the proposed model in IoT. Other related models (such as speech recognition, dialogue generation, etc.) are omitted due to space limitation.

IV. Experiments

In this section, some listening experiments to investigate the performance of our proposed multistage cross-lingual transfer learning scheme are conducted. The speech samples are available at the link: https://ttslr.github.io/IOT/.

A. Experimental Corpora

Large-Scale Neutral English TTS Corpus: The LJ-Speech database [67] consists of 13 100 short clips with a total of nearly 24 h of speech from one single speaker by reading seven nonfiction books.

Medium-Scale Emotional English TTS Corpus: A high-quality English emotional speech corpus[2] is recorded with 16-kHz sampling rate and 16 bits quantization in a professional studio. It involves 350 parallel English sentences performed by a female speaker in three different emotions (namely, neutral, happy, and sad), resulting in 1050 English utterances in total. The total duration of the emotional English TTS corpus is about 45 min, including 15, 12, and 18 min for neutral, happy, and sad emotions, respectively. There are about 6.31 words per English utterance on average.

Small-Scale Emotional Mandarin TTS Corpus: We record the emotional Mandarin TTS corpus in the same condition as the English one. For emotional Mandarin TTS corpus, 350 parallel utterances are performed by another female speaker also with 16-kHz sampling rate and 16 bits quantization in three different emotions (neutral, happy, and sad), leading to a total of 1050 Mandarin utterances. There are 18, 16, and 28 min for neutral, happy, and sad emotions, resulting in 62 min in total. What is more, there are 11.5 characters on average in each Mandarin utterance.

B. Experimental setup

To verify the effectiveness of the proposed three-stage deep transfer learning scheme, denoted as *3DTL-ETTS*, where three baseline systems (viz., ETTS, TL-ETTS, and 2DTL-ETTS) are compared as follows.

[2]https://github.com/HLTSingapore/Emotional-Speech-Data

TABLE II
HYPERPARAMETERS AND NETWORK ARCHITECTURE OF THE EMOTIONAL MANDARIN TTS MODEL, WHICH ARE SIMILAR WITH THE TACOTRON-BASED MODEL [20]

Model component	Hyper-parameters
Feature extraction	pre-emphasis: 0.97; frame length: 50-ms; frame shift: 12.5 ms; window type: Hann;
Character embedding	256 dimensional
Text Encoder	3 CNN layers: 512 filters with shape 5 × 1; 1 bi-LSTM layer: 512 units;
Emotion Encoder	FC-3-ReLU → Dropout(0.5) → FC-2-ReLU → Dropout(0.5);
Emotion vector	2 dimensional
Attention mechanism	Location sensitive attention
Decoder	FC-256-ReLU → Dropout(0.5) → FC-128-ReLU → Dropout(0.5); 2-layer LSTM: 512 units; 5 CNN layers: 512 filters with shape 5 × 1;

1) *ETTS [32]:* It is a one-stage training scheme [32] without using transfer learning technology. It is the Tacotron-based model architecture [33] introduced in Section II-B. We train the emotional Mandarin TTS model with 55-min data and 100k steps.
2) *TL-ETTS [40]:* It is a two-stage transfer learning or simple model adaptation scheme [40]. We first train the Mandarin neutral TTS model using the large-scale DataBaker Mandarin neutral TTS corpus[3] (12 h) as a prior. Then, we fine-tune the emotional Mandarin TTS model with the small-scale emotional Mandarin TTS corpus. After 100k steps pretraining, 15k steps are utilized to fine-tune the emotional Mandarin TTS model as the final model. Different with system ETTS, three emotion labels to meet the requirement of emotional Mandarin TTS corpus are employed to fine-tune the corresponding model.
3) *2DTL-ETTS:* It is a two-stage deep transfer learning scheme. Similar with system TL-ETTS using two-stage transfer learning, the difference between systems TL-ETTS and 2DTL-ETTS is that system 2DTL-ETTS takes advantage of different TTS corpus to do training. That is to say, the neutral English and emotional Mandarin TTS models are orderly trained and fine-tuned via English neutral and emotional Mandarin TTS corpora, respectively. Both of two stages take 100k steps, and more steps are required to fine-tune for better transferring knowledge compared to system TL-ETTS.
4) *3DTL-ETTS*: It is the proposed three-stage deep transfer learning scheme in Fig. 3. The same with system TL-ETTS, 100k steps are exploited to train our system, however, only 15k steps are used for the fine-tuning process, which is vastly less than that in system 2DTL-ETTS. Different from the above systems, the necessary stage-III, named self-adaptation stage, is added and trained with 100k steps.

The emotion vector is a 2-D continuous real vector, and a narrow emotion vector is also included to allow itself to focus on more meticulous emotional patterns. For Chinese experiments, the input of text encoder is pinyin sequence with tones, while it takes the character sequence as input for those English experiments. All systems are fed with 256-dimensional character sequence for their text encoders and both decoders output 80-channel mel-spectrum. Note that the decoder only generates one nonoverlapping output frame per each decoding step [20]. Additionally, following the Tacotron-based model [20], Adam optimizer [69] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is exploited and the size of batch is 32. We partition these 350 parallel utterances from each emotion into two different sets, namely, training set (first 330 utterances, 55 min) and test set (the rest 20 utterances, 7 min). The hyperparameters and setup of the network architecture, which followed by the Tacotron-based model [20], are tabulated in Table II. In this work, the chosen training steps are determined by the attention alignment performance. Model checkpoints were saved every 2000 steps and the best model's parameters are selected based on the performance of the validation set.

[3]https://www.data-baker.com/open_source.html

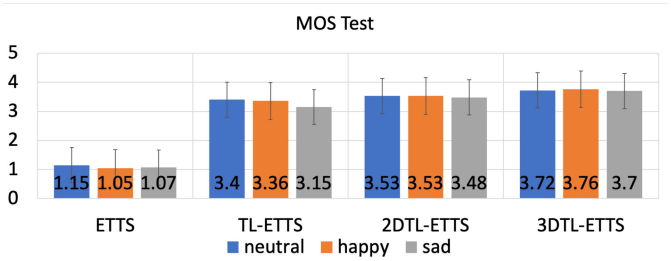

Fig. 5. MOS results of different systems, with 95% confidence interval computed by the *t*-test [68].

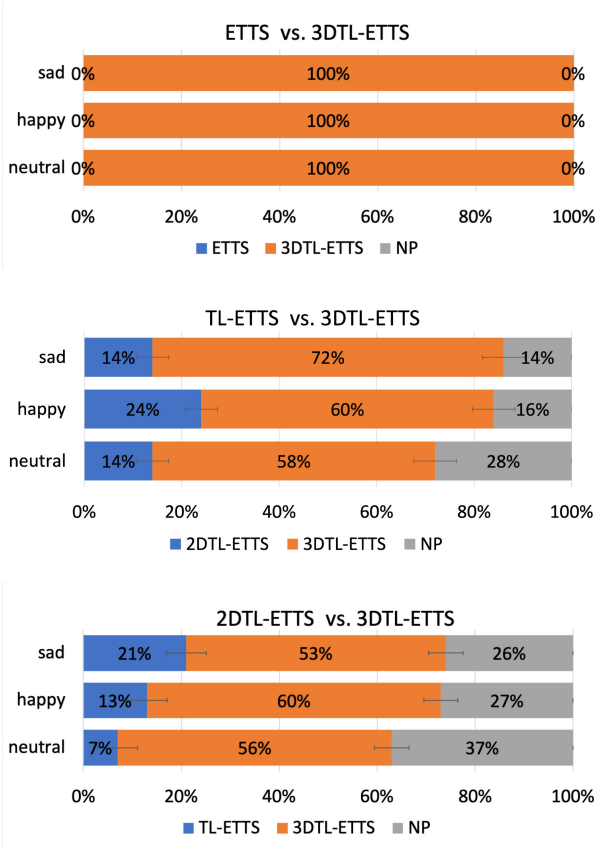

Fig. 6. AB preference results among compared systems, with 95% confidence interval computed by the *t*-test [68].

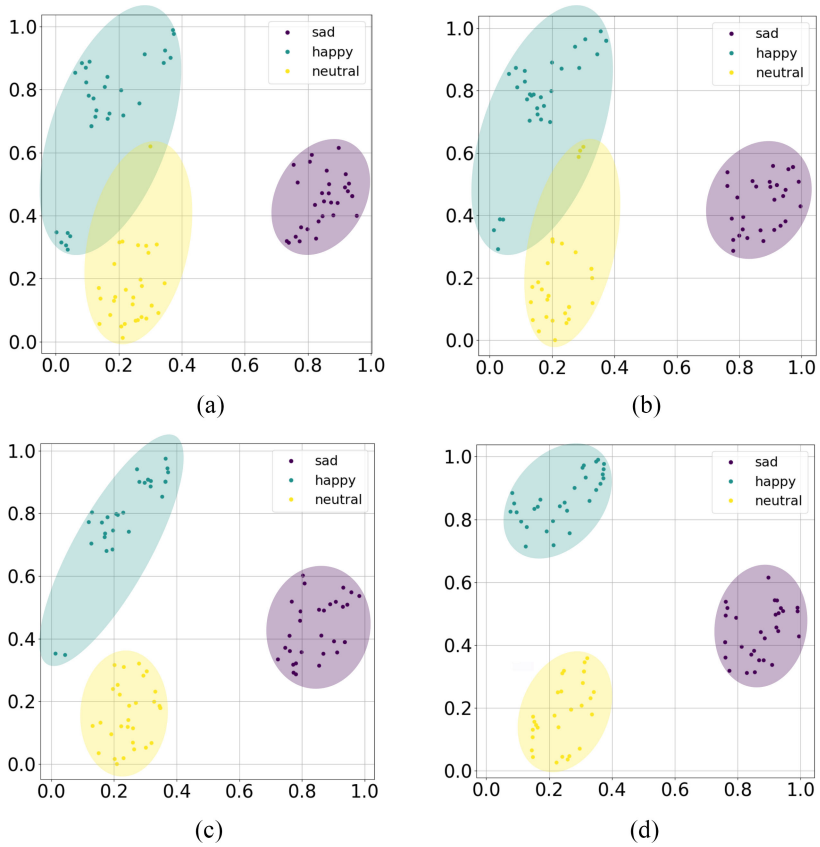

Fig. 7. Visualizations of the emotion vector for 30 utterances. (a) ETTS. (b) TL-ETTS. (c) 2DTL-ETTS. (d) 3DTL-ETTS.

For all experiments, Griffin–Lim algorithm [54] is applied to achieve rapid turn around for waveform generation. It is straightforward to replace Griffin–Lim by a WaveNet-like neural vocoder [27], [28], [55] to improve the audio fidelity.

C. Subjective Evaluation

To start with, a mean opinion score (MOS) test is implemented to evaluate the sound quality of the synthesized speech samples, and then followed by AB preference tests in terms of emotional expressiveness.

1) MOS Test for Naturalness: As presented in Fig. 5, the sound qualities of the synthesized speech samples are tested by MOS among ETTS, TL-ETTS, 2DTL-ETTS, and the proposed 3DTL-ETTS systems. The grading rule for listening test is set at a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. The MOS values are calculated by taking the arithmetic average of all scores assigned the subjects who have passed the validation question test. To robust against other interference factors, the linguistic contents are kept the same among all compared models. Here, 14 speakers (ten mandarin and four non-Native Mandarin speakers) are invited to participate the listening tests with 280 synthesized speech samples. From Fig. 5, we conclude the following.

1) *Importance of Transfer Learning:* Among the comparison between systems ETTS and TL-ETTS, it is observed that TL-ETTS performs better than ETTS for all emotions. The reason is because the pretraining technology can determine a suitable basis for model parameters and further facilitate optimizing and/or fine-tuning.
2) *Importance of Deep Transfer Learning:* Among the comparison between systems TL-ETTS, 2DTL-ETTS, and 3DTL-ETTS, we can see that systems 2DTL-ETTS and 3DTL-ETTS outperforms TL-ETTS, and the proposed system 3DTL-ETTS achieves the best result. To some extent, it is due to that the multistage cross-lingual transfer learning scheme successfully transfers the acoustic knowledge from English corpus to Mandarin emotional speech generation task.
3) *Importance of Three-Stage Deep Transfer Learning:* Among the comparison between systems TL-ETTS, 2DTL-ETTS, and 3DTL-ETTS, we can see that systems 2DTL-ETTS and 3DTL-ETTS outperforms TL-ETTS, and the proposed system 3DTL-ETTS achieves the best result. To some extent, it is due to that the multistage cross-lingual transfer learning scheme successfully transfers the acoustic knowledge from English corpus to the Mandarin emotional speech generation task.

2) AB Preference Test for Emotional Expressiveness: The AB preference tests are conducted to show the emotional expressiveness of the proposed system, as reported in Fig. 6. In AB preference tests, the listeners are asked to compare the quality and naturalness of the synthesized speech samples from different systems, and select the better one. From the results, we observe that listeners come to an agreement on our system showing more effective in emotional expressiveness, which is consistent with our previous analysis.

D. Visualization

As compared to the baseline systems, the proposed emotional Mandarin TTS system can obtain more natural and emotional speech due to the three-stage deep transfer learning scheme. To evaluate that the trained emotion vector in our proposed method can express emotion attitudes very well, we also use the t-SNE algorithm [70] to project the trained emotion vector into the 2-D space for all baseline systems and our 3DTL-ETTS system.

We randomly choose 30 utterances from the test set and extract the trained emotion vectors of all four systems. To sum up, for each system, there are 30 points for each emotion and 90 points in total. As shown in Fig. 7, we can see that all emotion vectors of the three emotion categories, which are sad, neural, and happy, are clearly clustered into three clusters for each system. Furthermore, we highlight the distributions of clusters in different colors. For our 3DTL-ETTS, the utterances within the same group form a cluster, while the utterances between groups are obviously separated from each other. Note that there is no obvious clustering boundary for the baselines, which means that our 3DTL-ETTS shows a better clustering than ETTS, TL-ETTS, and 2DTL-ETTS, and also demonstrates the proposed multistage deep transfer learning scheme has taken effect in robustness.

V. Conclusion

A novel training strategy for the emotional Mandarin TTS system is proposed, which includes three-stage cross-lingual transfer learning to tackle the emotional resource limitation problem. Herein, we implement a fine-grained emotion knowledge transfer process from English to Mandarin language. Some experimental results have been obtained to verify that the proposed three-stage cross-lingual transfer learning scheme outperforms all baseline systems in terms of the naturalness and emotion expressiveness. It is envisioned that the proposed system can make some progress in HCI, and is also attractive to further develop proper models for ETTS generation via IoT enablement.

References

[1] S. Chen, H. Xu, D. Liu, B. Hu, and H. Wang, "A vision of IoT: Applications, challenges, and opportunities with China perspective," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 349–359, Aug. 2014.

[2] Z. Gao, H.-Z. Xuan, H. Zhang, S. Wan, and K.-K. R. Choo, "Adaptive fusion and category-level dictionary learning model for multiview human action recognition," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9280–9293, Dec. 2019.

[3] S. Wu *et al.*, "Survey on prediction algorithms in smart homes," *IEEE Internet Things J.*, vol. 4, no. 3, pp. 636–644, Jun. 2017.

[4] H. Zhu *et al.*, "Smart healthcare in the era of Internet-of-Things," *IEEE Consum. Electron. Mag.*, vol. 8, no. 5, pp. 26–30, Sep. 2019.

[5] S. Feng, P. Setoodeh, and S. Haykin, "Smart home: Cognitive interactive people-centric Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 34–39, Feb. 2017.

[6] S. Li, L. Da Xu, and S. Zhao, "The Internet of Things: A survey," *Inf. Syst. Front.*, vol. 17, no. 2, pp. 243–259, 2015.

[7] M. Nitti, V. Pilloni, and L. Atzori, "EmIoT: Giving emotional intelligence to the Internet of Things," in *Proc. 10th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, 2018, pp. 1–3.

[8] H. Jeon, H. R. Oh, I. Hwang, and J. Kim, "An intelligent dialogue agent for the IoT home," in *Proc. Workshops 30th AAAI Conf. Artif. Intell.*, 2016, pp. 35–40.

[9] P. Ni, Y. Li, G. Li, and V. Chang, "Natural language understanding approaches based on joint task of intent detection and slot filling for IoT voice interaction," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 16149–16166, 2020.

[10] M. Johnston *et al.*, "MVA: The multimodal virtual assistant," in *Proc. 15th Annu. Meeting Special Interest Group Discourse Dial. (SIGDIAL)*, 2014, pp. 257–259.

[11] A. Pyae and T. N. Joelsson, "Investigating the usability and user experiences of voice user interface: A case of google home smart speaker," in *Proc. 20th Int. Conf. Human-Comput. Interact. Mobile Devices Services Adjunct*, 2018, pp. 127–131.

[12] B. A. Johnsson and B. Magnusson, "Towards end-user development of graphical user interfaces for Internet of Things," *Future Gener. Comput. Syst.*, vol. 107, pp. 670–680, Jun. 2020.

[13] T. Zhu and F. Zhang, "Design of marine two-way voice communication system based on human-computer interaction," *J. Coastal Res.*, vol. 95, pp. 1389–1394, May 2020.

[14] T. Iio, Y. Yoshikawa, M. Chiba, T. Asami, Y. Isoda, and H. Ishiguro, "Twin-robot dialogue system with robustness against speech recognition failure in human-robot dialogue with elderly people," *Appl. Sci.*, vol. 10, no. 4, p. 1522, 2020.

[15] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

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

[17] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Conf.*, vol. 1, 1996, pp. 373–376.

[18] T. Merritt, R. A. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2016, pp. 5145–5149.

[19] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[20] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 4779–4783.

[21] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4702.

[22] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 6940–6944.

[23] Y. Wang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.

[24] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2018, pp. 595–602.

[25] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust Tacotron-based TTS," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 6274–6278.

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

[27] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 1–15.

[28] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[29] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of wavenet as a statistical vocoder," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2018, pp. 5674–5678.

[30] A. Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3918–3926.

[31] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 6950–6954.

[32] Y. Lee, S.-Y. Lee, and A. Rabiee, "Emotional end-to-end neural speech synthesizer," in *Neural Information Processing Systems Foundation*. Red Hook, NY, USA: Curran Assoc.,2017.

[33] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1383–1387, Sep. 2019.

[34] N. Tits, K. El Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional tts," in *Proc. SAI Intell. Syst. Conf.*, 2019, pp. 52–60.

[35] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 7254–7258.

[36] R. Liu, B. Sisman, and H. Li, "Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability," in *Proc. Interspeech 2021*, 2021, pp. 1–5.

[37] G. Hofer, K. Richmond, and R. A. J. Clark, "Informed blending of databases for emotional speech synthesis," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 501–504.

[38] E. Navas, I. Hernáez, and I. Luengo, "An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1117–1127, Jul. 2006.

[39] "China to Become World's Largest IoT Market in 2024: Report." Global.Chinadaily.com. 2021. [Online]. Available: http://global.chinadaily.com.cn/a/202101/17/WS6003e0a3a31024ad0baa31af.html

[40] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2019, pp. 623–627.

[41] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, 2019, pp. 192–199.

[42] B. Joshi *et al.* "An Exploration into Deep Learning Methods for Emotional Text-to-Speech." Jun. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3876081

[43] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[44] W. F. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," *Semiotica*, vol. 158, no. 158, pp. 407–424, 2006.

[45] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *J. Phonetics*, vol. 37, no. 4, pp. 417–435, 2009.

[46] A.-J. Li, Y. Jia, Q. Fang, and J.-W. Dang, "Emotional intonation modeling: A cross-language study on Chinese and Japanese," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–6.

[47] P. Chhikara, P. Singh, R. Tekchandani, N. Kumar, and M. Guizani, "Federated learning meets human emotions: A decentralized framework for human–computer interaction for IoT applications," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6949–6962, Apr. 2021.

[48] L. Y. Mano *et al.*, "Exploiting IoT technologies for enhancing health smart homes through patient identification and emotion recognition," *Comput. Commun.*, vols. 89–90, pp. 178–190, Sep. 2016.

[49] M. Awais *et al.*, "LSTM based emotion detection using physiological signals: IoT framework for healthcare and distance learning in Covid-19," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16863–16871, Dec. 2021.

[50] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synth. Workshop*, 2002, pp. 227–230.

[51] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003, pp. 2461–2464.

[52] F. Eyben *et al.*, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2012, pp. 4009–4012.

[53] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Commun.*, vol. 99, pp. 135–143, May 2018.

[54] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[55] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, 2017, pp. 712–718.

[56] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[57] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, 2015, pp. 1225–1237.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[60] D. Soekhoe, P. Van Der Putten, and A. Plaat, "On the impact of data set size in transfer learning using deep neural networks," in *Proc. Int. Symp. intell. Data Anal.*, 2016, pp. 50–60.

[61] C. Wang, J. Pino, and J. Gu, "Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation," in *Proc. Interspeech*, 2020, pp. 4731–4735.

[62] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Nov. 2016, pp. 1568–1575. [Online]. Available: https://www.aclweb.org/anthology/D16-1163

[63] Y.-H. Lin *et al.*, "Choosing transfer languages for cross-lingual learning," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist.*, 2019, pp. 3125–3135.

[64] X. Chen, A. Hassan, H. Hassan, W. Wang, and C. Cardie, "Multi-source cross-lingual model transfer: Learning what to share," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist.*, 2019, pp. 3098–3112.

[65] Z. Li *et al.*, "Learn to cross-lingual transfer with meta graph learning across heterogeneous languages," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2020, pp. 2290–2301.

[66] C.-H. Lee and H.-Y. Lee, "Cross-lingual transfer learning for question answering," 2019, *arXiv:1907.06042*.

[67] K. Ito, 2017, "The LJ Speech Dataset," [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[68] G. E. Forsythe, *Computer Methods for Mathematical Computations* (Series in Automatic Computation), vol. 259. Englewood Cliffs, NJ, USA: Prentice-Hall, 1977.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[70] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.