applied sciences

MDPI

Review

A Review of Deep Learning Based Speech Synthesis

Yishuang Ning [1,2,3,4], Sheng He [1,2,3,4], Zhiyong Wu [5,6,*], Chunxiao Xing [1,2] and Liang-Jie Zhang [3,4]

1 Research Institute of Information Technology Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
2 Department of Computer Science and Technology Institute of Internet Industry, Tsinghua University, Beijing 100084, China
3 National Engineering Research Center for Supporting Software of Enterprise Internet Services, Shenzhen 518057, China
4 Kingdee Research, Kingdee International Software Group Company Limited, Shenzhen 518057, China
5 Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen Key Laboratory of Information Science and Technology, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
6 Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
* Correspondence: zywu@sz.tsinghua.edu.cn

Received: 1 August 2019; Accepted: 20 September 2019; Published: 27 September 2019

Abstract: Speech synthesis, also known as text-to-speech (TTS), has attracted increasingly more attention. Recent advances on speech synthesis are overwhelmingly contributed by deep learning or even end-to-end techniques which have been utilized to enhance a wide range of application scenarios such as intelligent speech interaction, chatbot or conversational artificial intelligence (AI). For speech synthesis, deep learning based techniques can leverage a large scale of <text, speech> pairs to learn effective feature representations to bridge the gap between text and speech, thus better characterizing the properties of events. To better understand the research dynamics in the speech synthesis field, this paper firstly introduces the traditional speech synthesis methods and highlights the importance of the acoustic modeling from the composition of the statistical parametric speech synthesis (SPSS) system. It then gives an overview of the advances on deep learning based speech synthesis, including the end-to-end approaches which have achieved start-of-the-art performance in recent years. Finally, it discusses the problems of the deep learning methods for speech synthesis, and also points out some appealing research directions that can bring the speech synthesis research into a new frontier.

Keywords: deep learning; speech synthesis; end-to-end; text analysis

1. Introduction

Speech synthesis, more specifically known as text-to-speech (TTS), is a comprehensive technology that involves many disciplines such as acoustics, linguistics, digital signal processing and statistics. The main task is to convert text input into speech output. With the development of speech synthesis technologies, from the previous formant based parametric synthesis [1,2], waveform concatenation based methods [3–5] to the current statistical parametric speech synthesis (SPSS) [6], the intelligibility and naturalness of the synthesized speech have been improved greatly. However, there is still a long way to go before computers can generate natural speech with high naturalness and expressiveness like that produced by human beings. The main reason is that the existing methods are based on shallow models that contain only one-layer nonlinear transformation units, such as hidden Markov models (HMMs) [7,8] and maximum Entropy (MaxEnt) [9]. Related studies show that shallow models

have good performance on data with less complicated internal structures and weak constraints. However, when dealing with the data having complex internal structures in the real world (e.g., speech, natural language, image, video, etc.), the representation capability of shallow models will be restricted.

Deep learning (DL) is a new research direction in the machine learning area in recent years. It can effectively capture the hidden internal structures of data and use more powerful modeling capabilities to characterize the data [10]. DL-based models have gained significant progress in many fields such as handwriting recognition [11], machine translation [12], speech recognition [13] and speech synthesis [14]. To address the problems existing in speech synthesis, many researchers have also proposed the DL-based solutions and achieved great improvements. Therefore, to summarize the DL-based speech synthesis methods at this stage will help us to clarify the current research trends in this area. The rest of the article is organized as follows. Section 2 gives an overview of speech synthesis including its basic concept, history and technologies. In Section 3, this paper introduces the pipeline of SPSS. A brief introduction is given in Section 4 about the DL-based speech synthesis methods including the end-to-end ones. Finally, Section 5 provides discussions on new research directions. Finally, Section 6 concludes the article.

2. An Overview of Speech Synthesis

2.1. Basic Concept of Speech Synthesis

Speech synthesis or TTS is to convert any text information into standard and smooth speech in real time. It involves many disciplines such as acoustics, linguistics, digital signal processing, computer science, etc. It is a cutting-edge technology in the field of information processing [15], especially for the current intelligent speech interaction systems.

2.2. The History of Speech Synthesis

With the development of digital signal processing technologies, the research goal of speech synthesis has been evolving from intelligibility and clarity to naturalness and expressiveness. Intelligibility describes the clarity of the synthesized speech, while naturalness refers to ease of listening and global stylistic consistency [16].

In the development of speech synthesis technology, early attempts mainly used parametric synthesis methods. In 1971, the Hungarian scientist Wolfgang von Kempelen used a series of delicate bellows, springs, bagpipes and resonance boxes to create a machine that can synthesize simple words. However, the intelligibility of the synthesized speech is very poor. To address this problem, in 1980, Klatt's serial/parallel formant synthesizer [17] was introduced. The most representative one is the DECtalk text-to-speech system of the Digital Equipment Corporation (DEC) (Maynard, MA, USA). The system can be connected to a computer through a standard interface or separately connected to the telephone network to provide a variety of speech services that can be understood by users. However, since the extraction of the formant parameters is still a challenging problem, the quality of the synthesized speech makes it difficult to meet the practical demand. In 1990, the Pitch Synchronous OverLap Add (PSOLA) [18] algorithm greatly improved the quality and naturalness of the speech generated by the time-domain waveform concatenation synthesis methods. However, since PSOLA requires the pitch period or starting point to be annotated accurately, the error of the two factors will affect the quality of the synthesized speech greatly. Due to the inherent problem of this kind of method, the synthesized speech is still not as natural as human speech. To tackle the issue, people conducted in-depth research on speech synthesis technologies, and used SPSS models to improve the naturalness of the synthesized speech. Typical examples are HMM-based [19] and DL-based [20] synthesis methods. Extensive experimental results demonstrate that the synthesized speech of these models has been greatly improved in both speech quality and naturalness.

2.3. Traditional Speech Synthesis Technology

To understand why deep learning techniques are being used to generate speech today, it is important to know how speech generation is traditionally done. There are two specific methods for TTS conversion: concatenative TTS and parametric TTS. This paper will give a brief introduction to the two kinds of methods in the following sections.

2.3.1. Concatenative Speech Synthesis

The waveform concatenation based synthesis method directly concatenates the waveforms in the speech waveform database and outputs a continuous speech stream. Its basic principle is to select the appropriate speech unit from the pre-recorded and labeled speech corpus according to the context information analyzed from the text input, and concatenate the selected speech unit to obtain the final synthesized speech. With the guidance of the context infomation, the naturalness of the synthesized speech has been improved greatly.

There are two different schemes for concatenative synthesis: one is based on linear prediction coefficients (LPCs) [21], the other is based on PSOLA. The first method mainly uses the LPC coding of speech to reduce the storage capacity occupied by the speech signal, and the synthesis is also a simple decoding and concatenation process. The speech synthesized by this method is very natural for a single word because the codec preserves most of the information of the speech. However, since the natural flow of words when people actually speak is not just a simple concatenation of individual isolated speech units, the overall effect will be affected by the concatenative points. To address this problem, PSOLA, which pays more attention to the control and modification of prosody, has been proposed. Different from the former method, PSOLA adjusts the prosody of the concatenation unit according to the target context, so that the final synthesized waveform not only maintains the speech quality of the original pronunciation, but also makes the prosody features of the concatenation unit conform to the target context. However, this method also has many defects: (1) as stated in Section 2.2, the quality of the synthesized speech will be affected by the pitch period or starting point; and (2) the problem of whether it can maintain a smooth transition has not been solved. These defects greatly limit its application in diversified speech synthesis [22].

2.3.2. Parametric Speech Synthesis

The parametric speech synthesis refers to the method that uses digital signal processing technologies to synthesize speech from text. In this method, it considers the human vocal process as a simulation that uses a source of glottal state to excite a time-varying digital filter which characterizes the resonance characteristics of the channel. The source can be a periodic pulse sequence that is used to represent the vocal cord vibration of the voiced speech, or a random white noise to indicate undefined unvoiced speech. By adjusting the parameters of the filter, it can synthesize various types of speeches [15]. Typical methods include vocal organ parametric synthesis [23], formant parametric synthesis [24], HMM-based speech synthesis [25], and deep neural network (DNN)-based speech synthesis [26,27].

3. Statistical Parametric Speech Synthesis

A complete SPSS system is generally composed of three modules: a text analysis module, a parameter prediction module which uses a statistical model to predict the acoustic feature parameters such as fundamental frequency (F0), spectral parameters and duration, and a speech synthesis module. The text analysis module mainly preprocesses the input text and transforms it into linguistic features used by the speech synthesis system, including text normalization [28], automatic word segmentation [20], and grapheme-to-phoneme conversion [29]. These linguistic features usually include phoneme, syllable, word, phrase and sentence-level features. The purpose of the parameter prediction module is to predict the acoustic feature parameters of the target speech according to

the output of the text analysis module. The speech synthesis module generates the waveform of the target speech according to the output of the parameter prediction module by using a particular synthesis algorithm. The SPSS is usually divided into two phases: the training phase and the synthesis phase. In the training phase, acoustic feature parameters such as F0 and spectral parameters are firstly extracted from the corpus, and then a statistical acoustic model is trained based on the linguistic features of the text analysis module as well as the extracted acoustic feature parameters. In the synthesis phase, the acoustic feature parameters are predicted using the trained acoustic model with the guidance of the linguistic features. Finally, the speech is synthesized based on the predicted acoustic feature parameters using a vocoder.

3.1. Text Analysis

Text analysis is an important module of the SPSS model. Traditional text analysis methods are mainly rule-based, which require a lot of time to collect and learn these rules. With the rapid development of data mining technology, some data-driven methods have been gradually developed, such as the bigram method, trigram method, HMM-based method and DNN-based method. When using the latter two methods for text analysis, the Festival [4] system is usually used to perform phoneme segmentation and annotation on the corpus, which mainly includes the following five levels:

Phoneme level: the phonetic symbols of the previous before the previous, the previous, the current, the next or the next after the next; the forward or backward distance of the current phoneme within the syllable.

Syllable level: whether the previous, the current or the next syllable is stressed; the number of phonemes contained in the previous, the current or the next syllable; the forward or the backward distance of the current syllable within the word or phrase; the number of the stressed syllables before or after the current syllable within the phrase; the distance from the current syllable to the forward or backward most nearest stressed syllable; the vowel phonetics of the current syllable.

Word level: the part of speech (POS) of the previous, the current or the next word; the number of syllables of the previous, the current or the next word; the forward or backward position of the current word in the phrase; the forward or backward content word of the current word within the phrase; the distance from the current word to the forward or backward nearest content word; the POS of the previous, the current or the next word.

Phrase level: the number of syllables of the previous, the current or the next phrase; the number of words of the previous, the current or the next phrase; the forward or backward position of the current phrase in the sentence; the prosodic annotation of the current phrase.

Sentence level: The number of syllables, words or phrases in the current sentence.

3.2. Parameter Prediction

Parameter prediction is used to predict acoustic feature parameters based on the result of the text analysis module and the trained acoustic model. For the SPSS, there are usually two kinds of parameter prediction methods: HMM-based parameter prediction and DNN-based parameter prediction. This paper will give a review of these methods in the following.

3.2.1. HMM-Based Parameter Prediction

The HMM-based parameter prediction method mainly generates the sequence of F0 and spectral parameters from the trained HMMs. It is achieved by calculating the sequence of acoustic features with the maximum likelihood estimation (MLE) algorithm given a Gaussian distribution sequence. Due to the differences between F0 and spectral parameters, different methods have been adopted to model the two kinds of feature parameters. For the continuous spectral parameters, the continuous density hidden Markov model (CD-HMM) is used and the output of each HMM state is a single Gaussian or a Gaussian mixture model (GMM) [27]. However, for the variable-dimensional F0

parameters which include voiced and unvoiced regions, it is difficult to apply discrete or continuous HMMs because the values of F0 are not defined in unvoiced regions. To address this problem, the HMM-based method adopts multi-space probability distribution to model the voiced and unvoiced regions (e.g., voiced/unvoiced (V/UV) parameters), separately. To improve the accuracy and flexibility of acoustic parameter prediction, the authors in [28] introduce the articulatory feature that is related to the speech generation mechanism and integrates it with the acoustic features.

3.2.2. DNN-Based Parameter Prediction

It is well known that the acoustic features of a particular phoneme will be affected by the context information associated with the phoneme [30]. It means that the context information plays a significant role in the prediction of the acoustic features. Researchers show that the human speech generation process usually uses a hierarchical structure to convert the context information into a speech waveform [31]. Inspired by this idea, the deep structure models have been introduced in predicting acoustic feature parameters for speech synthesis [32]. The framework of the DNN-based parameter prediction progress can be seen in [20].

To compare with the HMM-based parameter prediction methods, the DNN-based methods can not only map complex linguistic features into acoustic feature parameters, but also use long short-term context information to model the correlation between frames which improves the quality of speech synthesis. In addition, for the HMM-based methods, the principal of MLE is used to maximize the output probability which makes the parameter sequence a mean vector sequence, resulting in a step-wise function. The jumps cause discontinuities in the synthesized speech. To address this problem, the maximum likelihood parameter generation (MLPG) algorithm is used to smooth the trajectory by taking the dynamic features including the delta and delta–delta coefficients into account. However, the DNN-based methods cannot suffer from this problem.

3.3. Vocoder-Based Speech Synthesis

Speech synthesizer or vocoder is an important component of statistical parametric speech synthesis, which aims at synthesizing speech waveform based on the estimated acoustic feature parameters. Traditional methods usually use the HTS_engine [33] synthesizer since it is free and fast to synthesize speech. However, the synthesized speech usually sounds dull, thus making the quality not good. To improve the quality of the synthesized speech, STRAIGHT [34,35] is proposed and used in various studies, making it easy to manipulate speech. Other methods such as phase vocoder [36], PSOLA [18] and sinusoidal model [37] are also proposed. Legacy-STRAIGHT [38] and TANDEM-STRAIGHT [38] were developed as algorithms to meet the requirements for high-quality speech synthesis. Although these methods can synthesize speech with good quality, the speed still cannot meet the real-world application scenarios. To address this problem, real-time methods remain a popular research topic. For example, the authors in [34] proposed the real-time STRAIGHT as a way to meet the demand for real-time processing. The authors in [38] proposed a high-quality speech synthesis system which used WORLD [39] to meet the requirements of not only high sound quality but also real-time processing.

4. Deep Learning Based Speech Synthesis

It is known that the HMM-based speech synthesis method maps linguistic features into probability densities of speech parameters with various decision trees. Different from the HMM-based method, the DL-based method directly perform mapping from linguistic features to acoustic features with deep neural networks which have proven extraordinarily efficient at learning inherent features of data. In the long tradition of studies that adopt DL-based method for speech synthesis, people have proposed numerous models. To help readers better understand the development process of these methods (Audio samples of different synthesis methods are given at: http://www.ai1000.org/sampl

es/index.html.), this paper gives a brief overview of the advantages and disadvantages in Table 1 and makes a detailed introduction in the following.

Table 1. The advantages and disadvantages of different speech synthesis methods, including hidden Markov model (HMM), restrictive Boltzmann machine (RBM), deep belief network (DBN), deep mixture density network (DMDN), deep bidirectional long short-term memory (DBLSTM), WaveNet, Tacotron and convolutional neural network (CNN).

Methods	Advantages	Disadvantages
HMM	Flexible with changing voice characteristics and the system is robust	The acoustic features are oversmoothed, making the generated speech sounds muffled
RBM	Can better describe the distribution of high-dimensional spectral envelopes to alleviate the over-smooth problem	Suffer from the fragementation problem of training data
DBN	Cannot suffer from the training data fragementation problem and reduce the over-smoothing problem	The quality of generated speech will be degraded
DMDN	Can solve the single modality problem	Can only leverage limited contexts and each frame is mapped independently
DBLSTM	Can fully leverage contextual information	Still needs a vocoder to synthesize waveform
WaveNet	Can produce high-quality speech waveforms	Too slow and the errors from the front-end will affect the synthesis effect
Tacotron	Fully end-to-end speech synthesis model and can produce high-quality speech waveforms	Quite costly to train the model
CNN	Fast to train the model	The speech quality might be degraded

4.1. Restrictive Boltzmann Machines for Speech Synthesis

In recent years, restricted Boltzmann machines (RBMs) [40] have been widely used for modeling speech signals, such as speech recognition, spectrogram coding and acoustic-articulatory inversion mapping [40]. In these applications, RBM is often used for pre-training of deep auto-encoders (DAEs) [41,42] or DNNs. In the field of speech synthesis, RBM is usually regarded as a density model for generating the spectral envelope of acoustic parameters. It is adopted to better describe the distribution of high-dimensional spectral envelopes to alleviate the over-smooth problem in HMM-based speech synthesis [40]. After training the HMMs, a state alignment is performed for the acoustic features and the state boundaries are used to collect the spectral envelopes obtained from each state. The parameters of the RBM are estimated using the maximum likelihood estimation (MLE) criterion. Finally, RBM–HMMs are constructed to model the spectral envelopes. In the synthesis phase, the optimal spectral envelope sequence is estimated based on the input sentence and the trained RBM–HMMs. Although the subjective evaluation result of this method is better than that of traditional HMM–GMM systems, and the predicted spectral envelope is closer to the original one, this method still cannot solve the fragementation problem of training data encountered in the traditional HMM-based method.

4.2. Multi-Distribution Deep Belief Networks for Speech Synthesis

The multi-distribution deep belief network (DBN) [43] is a method of modeling the joint distribution of context information and acoustic features. It models the coutinuous spectral, discrete voiced/unvoiced (V/UV) parameters and the multi-space F0 simultaneously with three types of RBMs. Due to the different data types of the 1-out-of-*K* code, the F0, the spectral and the V/UV parameters, the method uses the 1-out-of-*K* code of the syllable and its corresponding acoustic parameters as the visible-layer data of the RBM to train the RBMs. In DBNs, the visible unit can obey different probability distributions; therefore, it is possible to characterize the supervectors that are composed of these features. In the training phase, given the 1-out-of-*K* code of the syllable, the network fixes the visible-layer units to calculate the hidden-layer parameters firstly, and then uses the parameters of the hidden layers to calculate the visible-layer parameters until convergence. Finally, the predicted acoustic features are interpolated based on the length of the syllable.

The advantage of this method is that all the syllables are trained in the same network, and all the data are used to train the same RBM or DBN. Therefore, it cannot suffer from the training data fragementation problem. In addition, modeling the acoustic feature parameters of a syllable directly can describe the correlation of each frame of the syllable and the correlation of different dimensions of the same frame. The method avoids averaging the frames corresponding to the same syllable, thus reducing the over-smooth phenomenon. However, since this method does not distinguish syllables in different contexts, it still averages the acoustic parameters corresponding to the same syllable. In addition, compared to the high-dimensional spectral parameters, the one-dimensional F0s don't contribute much to the model, thus making the predicted F0s contain a lot of noise that reduces the quality of the synthesized speech.

4.3. Speech Synthesis Using Deep Mixture Density Networks

Although the DNN-based speech synthesis model can synthesize speech with high naturalness, it still has some limitations to model acoustic feature parameters, such as the single modality of the objective function and the inability to predict the variance. To address these problems, the authors in [44] proposed the parameter prediction method based on a deep mixture density network, which uses a mixture density output layer to predict the probability distribution of output features under given input features.

4.3.1. Mixture Density Networks

Mixed density networks (MDNs) [45] can not only map input features to GMM parameters (such as the mixture weights, mean and variance), but also give the joint probability density function of y given input features x. The joint probability density function is expressed as follows:

$$p(y|x, M) = \sum_{m=1}^{M} w_m(x) \cdot N\left(y; \mu_m(x), \sigma_m^2(x)\right), \tag{1}$$

where M is the number of mixture components, and $w_m(x)$, $\mu_m(x)$ and $\sigma_m^2(x)$ are the mixture weights, mean and variance of the m-th Gaussian component of GMM, respectively. The parameters of the GMM can be calculated based on MDN with Equations (2)–(4):

$$w_m(x) = \frac{exp\left(z_m^{(w)}(x, M)\right)}{\sum_{l=1}^{M} exp\left(z_l^{(w)}(x, M)\right)}, \tag{2}$$

$$\sigma_m(x) = exp\left(z_m^{(\sigma)}(x, M)\right), \tag{3}$$

$$\mu_m(x) = z_m^{(\mu)}(x, M), \tag{4}$$

where $z_m^{(w)}(x, M)$, $z_m^{(\mu)}(x, M)$ and $z_m^{(\sigma)}(x, M)$ are the excitation of the MDN output layer corresponding to the mixture weights, mean and variance of the m-th Gaussian component, respectively. Finally, given the input/output pair of the training data in Equation (5), the model is trained by maximizing the log likelihood of M, which is expressed as Equation (6):

$$D = \left\{\left(x_1^{(1)}, y_1^{(1)}\right), ..., \left(x_{T(1)}^{(1)}, y_{T(1)}^{(1)}\right), ..., \left(x_1^{(N)}, y_1^{(N)}\right), ..., \left(x_{T(N)}^{(N)}, y_{T(N)}^{(N)}\right)\right\}, \tag{5}$$

$$\hat{M} = \arg\max_{M} \sum_{n=1}^{N} \sum_{t=1}^{T(n)} logp\left(y_t^{(n)} | x_t^{(n)}, M\right), \tag{6}$$

where N is the number of sentences and $T(n)$ is the number of frames in the nth sentence.

4.3.2. Deep MDN-Based Speech Synthesis

When predicting speech parameters with deep MDN, the text prompt is first converted into a linguistic feature sequence $\{x_1, x_2, ..., x_T\}$, and then the duration of each speech unit is predicted using a duration prediction model. The acoustic features including the F0, spectral parameters and their corresponding dynamic features are estimated with the forward algorithm and the trained deep MDN. Finally, the acoustic feature parameters are generated by the parameter generation algorithm and speech is synthesized with a vocoder.

4.4. Deep Bidirectional LSTM-Based Speech Synthesis

Although the deep MDN speech synthesis model can solve the single modality problem of the objective function and predict the acoustic feature parameters accurately to improve the naturalness of the synthesized speech, there are still some problems as elaborated in the following. Firstly, MDN can only leverage limited contextual information since it can only model fixed time span (e.g., fixed number of preceding or succeeding contexts) for input features. Secondly, the model can only do frame-by-frame mapping (e.g., each frame is mapped independently). To address these problems, the authors in [46] proposed a modeling method based on recurrent neural networks (RNNs). The advantage of RNN is the ability to utilize context information when mapping inputs to outputs. However, traditional RNNs can only access limited context information since the effects of a given input on the hidden layer and the output layer will decay or explode as it propagates through the network. In addition, this algorithm also cannot learn long-term dependencies.

To address these problems, the authors in [47] introduced a memory cell and proposed the long short-term memory (LSTM) model. To fully leverage contextual information, bidirectional LSTM [48] is mostly used for mapping the input linguistic features to acoustic features.

4.4.1. BLSTM

BLSTM-RNN is an extended architecture of bidirectional recurrent neural network (BRNN) [49]. It replaces units in the hidden layers of BRNN with LSTM memory blocks. With these memory blocks, BLSTM can store information for long and short time lags, and leverage relevant contextual dependencies from both forward and backward directions for machine learning tasks. With a forward and a backward layer, BLSTM can utilize both the past and future information for modeling.

Given an input sequence $x = (x_1, x_2, ..., x_T)$, BLSTM computes the forward hidden sequence $\overrightarrow{h}$ and the backward hidden sequence $\overleftarrow{h}$ by iterating the forward layer from $t = 1$ to T and the backward layer from $t = T$ to 1:

$$\overrightarrow{h}_t = \phi\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right), \tag{7}$$

$$\overleftarrow{h}_t = \phi\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right). \tag{8}$$

The output layer is connected to both forward and backward layers, thus the output sequence can be written as:

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y. \tag{9}$$

The notations of these equations are explained in [49] and $\phi(\cdot)$ is the activation function which can be implemented by the LSTM block with equations in [49].

4.4.2. Deep BLSTM-Based Speech Synthesis

When using a deep BLSTM-based (DBLSTM) model to predict acoustic parameters, first we need to convert the input text prompt into a feature vector, and then use the DBLSTM model to map the

input feature to acoustic parameters. Finally, the parameter generation algorithm is used to generate the acoustic parameters and a vocoder is utilized to synthesize the corresponding speech. For instance, the authors in [48] proposed a multi-task learning [50,51] of structured output layer (SOL) BLSTM model for speech synthesis, which is capable of balancing the error cost functions associated with spectral feature and pitch parameter targets.

4.5. Sequence-to-Sequence Speech Synthesis

Sequence-to-sequence (seq2seq) neural networks can transduce an input sequence into an output sequence that may have a different length and have been applied to various tasks such as machine translation [52], speech recognition [53] and image caption generation [54], and achieved promising results. Since speech synthesis is the reverse process of speech recognition, the seq2seq modeling technique has also been applied to speech synthesis recently. For example, the authors in [55] employed the structure with content-based attention [56] to model the acoustic features for speech synthesis. Char2Wav [16] adopted location-based attention to build an encoder–decoder acoustic model. To tackle the instability problem of missing or repeating phones that current seq2seq models still suffer from, the authors in [57] proposed a forward attention approach for the seq2seq acoustic modeling of speech synthesis. Tacotron, which is also a seq2seq model with an attention mechanism, has been proposed to map the input text to mel-spectrogram for speech synthesis.

4.6. End-to-End Speech Synthesis

A TTS system typically consists of a text analysis front-end, an acoustic model and a speech synthesizer. Since these components are trained independently and rely on extensive domain expertise which are laborious, errors from each component may compound. To address these problems, end-to-end speech synthesis methods which combine those components into a unified framework have become mainstream in the speech synthesis field. There are many advantages of an end-to-end TTS system: (1) it can be trained based on a large scale of <text, speech> pairs with minimum human annotation; (2) it doesn't require phoneme-level alignment; and (3) errors cannot compound since it is a single model. In the following, we will give a brief introduction to the end-to-end speech synthesis methods.

4.6.1. Speech Synthesis Based on WaveNet

WaveNet [58], which is evolved from the PixelCNN [59] or PixelRNN [60] model applied in image generation field, is a powerful generative model of raw audio waveforms. It was proposed by Deepmind (London, UK) in 2016 and opens the door for end-to-end speech synthesis. It is capable of generating relatively realistic-sounding human-like voices by directly modeling waveforms using a DNN model which is trained with recordings of real speech. It is a complete probabilistic autoregressive model that predicts the probability distribution of the current audio sample based on all samples that have been generated before. As an important component of WaveNet, dilated causal convolutions are used to ensure that WaveNet can only use the sampling points from 0 to $t-1$ while generating the tth sampling point.

The original WaveNet model uses autoregressive connections to synthesize waveforms one sample at a time, with each new sample conditioned on the previous ones. The joint probability of a waveform $X = \{x_1, x_2, ..., x_T\}$ can be factorised as follows:

$$p(X) = \prod_{i=0}^{T-1} p(x_{i+1}|x_1, x_2, ..., x_i). \tag{10}$$

Like other speech synthesis models, WaveNet-based models can be divided into training phase and generation phase. At the training phase, the input sequences are real waveforms recorded from human speakers. At the generation phase, the network is sampled to generate synthetic utterances.

To generate speech of the specified speaker or the specified text, global and local conditions are usually introduced to control the synthesis contents.

While the WaveNet model can produce high-quality audios, it still suffers from the following problems: (1) it is too slow because the prediction of each sampling point always depends on the predicted sampling points before; (2) it also depends on linguistic features from an existing TTS front-end and the errors from the front-end text analysis will directly affect the synthesis effect.

To address these problems, the parallel WaveNet is proposed to improve the sampling efficiency. It is capable of generating high-fidelity speech samples at more than 20 times faster [61]. Another neural model, Deep Voice [62], is also proposed to replace each component including a text analysis front-end, an acoustic model and a speech synthesizer by a corresponding neural network. However, since each component is trained independently, it is not a real end-to-end synthesis.

4.6.2. Speech Synthesis Based on Tacotron

Tacotron [63,64] is a fully end-to-end speech synthesis model. It is capable of training a speech synthesis model given <text, audio> pairs, thus alleviating the need for laborious feature engineering. In addition, since it is based on character level, it can be applied in almost all kinds of languages including Chinese Mandarin.

Like WaveNet, the Tacotron model is also a generative model. Different from WaveNet, Tacotron uses a seq2seq model with an attention mechanism to map text to a spectrogram, which is a good representation of speech. Since a spectrogram doesn't contain phase information, the system uses the Griffin–Lim algorithm [65] to reconstruct the audio by estimating the phase information from the spectrogram iteratively. The overall framework of the Tacotron speech synthesis model can be seen in [63].

Since Tacotron is a fully end-to-end model that directly maps the input text to mel-spectrogram, it has received a wide amount of attention of researchers and various improved versions have been proposed. For example, some researchers implemented open clones of Tacotron [66–68] to reproduce the speech of satisfactory quality as clear as the original work [69]. The authors in [70] introduced deep generative models, such as Variational Auto-encoder (VAE) [71], to Tacotron to explicitly model the latent representation of a speaker state in a continuous space, and additionally to control the speaking style in speech synthesis [70].

There are also some works that combine Tacotron and WaveNet for speech synthesis, such as Deep Voice 2 [72]. In this system, Tacotron is used to transform the input text to the linear scale spectrogram, while WaveNet is used to generate speech from the linear scale spectrogram output of Tacotron. In addition, the authors in [73] also proposed the Tacotron2 system to generate audio signals that resulted in a very high mean opinion score (MOS) comparable to human speech [74]. The authors in [73] described a unified neural approach that combines a seq2seq Tacotron-style model to generate mel-spectrogram and a WaveNet vocoder to synthesize speech from the generated mel-spectrogram.

4.6.3. Speech Synthesis Based on Convolutional Neural Networks (CNNs)

Although the Tacotron-based end-to-end system has achieved promising performance recently, it still has a drawback that there are many recurrent units. This kind of structure makes it quite costly to train the model and it is also infeasible for researchers without high-performance machines to conduct further research on it. To address this problem, a lot of works have been proposed. The authors in [69] proposed a deep convolutional network with guided attention which can be trained much faster than the RNN-based state-of-the-art neural system. Different from the WaveNet model, which utilized the fully-convolutional structure as a kind of vocoder or a back-end, Ref. [69] is rather a frond-end (and most of back-end processing) that can synthesize a spectrogram. The authors in [75] used CNN-based architecture for capturing long-term dependencies of singing voice and applied parallel computation to accelerate the model train and acoustic feature generation processes. The authors in [76] proposed a novel, fully-convolutional character-to-spectrogram architecture, namely Deep

Voice 3, for speech synthesis, which enables fully parallel computation to make the training process faster than that of using recurrent units.

5. Discussion

Compared with the concatenative speech synthesis method, the SPSS system can synthesize speech with high intelligibility and naturalness. Due to the limitations of the HMM-based speech synthesis model (such as the use of context decision trees to share speech parameters), the synthesized speech is not vivid enough to meet the requirements of expressive speech synthesis. The DL-based speech synthesis models adopt complete context information and distributed representation to replace the clustering process of the context decision tree in HMM, and use multiple hidden layers to map the context features to high-dimensional acoustic features, thus making the quality of the synthesized speech better than the traditional methods.

However, the powerful representation capabilities of DL-based models have also brought some new problems. To achieve better results, the models need more hidden layers and nodes, which will undoubtedly increase the number of parameters in the network, and the time complexity and space complexity for network training. When the training data are insufficient, the models usually have over-fitting. Therefore, it requires a large amount of corpora and computing resources to train the network. In addition, the DL-based models also require much more space to store the parameters.

There is no doubt that the existing end-to-end models are still far from perfect [77]. Despite many achievements, there are still some challenging problems. Next, we will discuss some research directions:

- Investigating context features hidden in end-to-end speech synthesis. The end-to-end TTS system, mostly back-end, has achieved state-of-the-art performance since it was proposed. However, there is little progress in front-end text analysis, which extracts context features or linguistic features that are very useful to bridge the gap between text and speech [78]. Therefore, demonstrating what types of context information are utilized in end-to-end speech synthesis system is a good direction in future.
- Semi-supervised or unsupervised training in end-to-end speech synthesis. Although end-to-end TTS models have shown excellent results, they typically require large amounts of high-quality <text, speech> pairs for training, which are expensive and time-consuming to collect. It is important and of great significance to improve the data efficiency for end-to-end TTS training by leveraging a large scale of publicly available unpaired text and speech recordings [79].
- The application of other speech related scenarios. In addition to the application of text-to-speech in this paper, the application to other scenarios such as voice conversion, audio-visual speech synthesis, speech translation and cross-lingual speech synthesis is also a good direction.
- The combination of software and hardware. At present, most deep neural networks require a lot of calculations. Therefore, parallelization will be an indispensable part of improving network efficiency. In general, there are two ways to implement parallelization: one is the parallelization of the machines; the other is to use GPU parallelization. However, since writing GPU code is still time-consuming and laborious for most researchers, it depends on the cooperation of hardware vendors and software vendors, to provide the industry with more and more intelligent programming tools.

6. Conclusions

Deep learning that is capable of leveraging large amount of training data has become an important technique for speech synthesis. Recently, increasingly more researches have been conducted on deep learning techniques or even end-to-end frameworks and achieved state-of-the-art performance. This paper gives an overview to the current advances on speech synthesis and compare both of the advantages and disadvantages among different methods, and discusses possible research directions that can promote the development of speech synthesis in the future.

Author Contributions: Conceptualization, Investigation, Writing—original draft, Writing—review and editing, Y.N.; Writing—review and editing, S.H.; Resources, Writing—review and editing and funding acquisition, Z.W.; Supervision and funding acquisition, C.X. and L.-J.Z.

Funding: This work is partially supported by the technical projects No. c1533411500138 and No. 2017YFB0802700. This work is also supported by the National Natural Science Foundation of China (NSFC) (91646202, 61433018), joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002), the 1000-Talent program and the China Postdoctoral Science Foundation (2019M652949).

Acknowledgments: The authors would like to thank Peng Liu, Quanjie Yu and Zhiyong Wu for providing the material.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

TTS	Text-to-Speech
SPSS	Statistical Parametric Speech Synthesis
HMM	Hidden Markov Model
DL	Deep Learning
DEC	Digital Equipment Corporation
POS	Part-of-Speech
DNN	Deep Neural Network
LPC	Linear Prediction Coefficient
PSOLA	Pitch Synchronous OverLap Add
SPSS	Statistical Parametric Speech Synthesis
CD-HMM	Continuous Density Hidden Markov Model
GMM	Gaussian Mixture Model
RBM	Restricted Boltzmann Machines
DBN	Deep Belief Networks
MLE	Maximum Likelihood Estimation
MDN	Mixed Density Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BLSTM	Bidirectional Long Short-term Memory
CNN	Convolutional Neural Network
DAE	Deep Auto-Encoder
VAE	Variational Auto-Encoder
MOS	Mean Opinion Score

References

1. Klatt, D.H. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.* **1987**, *82*, 737–793. [CrossRef] [PubMed]
2. Allen, J.; Hunnicutt, M.S.; Klatt, D.H.; Armstrong, R.C.; Pisoni, D.B. *From Text to Speech: The MITalk System*; Cambridge University Press: New York, NY, USA, 1987.
3. Murray, I.R.; Arnott, J.L.; Rohwer, E.A. Emotional stress in synthetic speech: Progress and future directions. *Speech Commun.* **1996**, *20*, 85–91. [CrossRef]
4. Festival. Available online: http://www.cstr.ed.ac.uk/projects/festival/ (accessed on 3 July 2019).
5. Chu, M.; Peng, H.; Zhao, Y. Microsoft Mulan. A bilingual TTS system. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), Hong Kong, China, 6–10 April 2003; pp. 264–267.
6. Tokuda, K.; Nankaku, Y.; Toda, T. Speech synthesis based on hidden Markov models. *Proc. IEEE* **2013**, *101*, 1234–1252. [CrossRef]
7. Murray, I.R. *Simulating Emotion in Synthetic Speech*; University of Dundee: Dundee, UK, 1989.
8. Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5–9 June 2000; Volume 3, pp. 1315–1318.

9. Ratnaparkhi, A. *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*; University of Pennsylvania: Philadelphia, PA, USA, 1997.
10. Yang, J.A.; Wang, Y.; Liu, H.; Li, J.H.; Lu, J. Deep learning theory and its application in speech recognition. *Commun. Countermeas.* **2014**, *33*, 1–5.
11. Graves, A. *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin, Germany, 2012.
12. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QB, Canada, 8–13 December 2014; pp. 3104–3112.
13. Graves, A.; Mohamed, A.R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
14. Zen, H.; Tokuda, K.; Alan, W.B. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [CrossRef]
15. Xu, S.H. *Study on HMM-Based Chinese Speech Synthesis*; Beijing University of Posts and Telecommunications: Beijing, China, 2007.
16. Sotelo, J.; Mehri, S.; Kumar, K.; Santos, J.F.; Kastner, K.; Courville, A.; Bengio, Y. Char2wav: End-to-end Speech Synthesis. In Proceedings of the International Conference on Learning Representations Workshop, Toulon, France, 24–26 April 2017.
17. Klatt, D.H. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* **1980**, *67*, 971–995. [CrossRef]
18. Moulines, E.; Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphone. *Speech Commun.* **1990**, *9*, 453–456. [CrossRef]
19. Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Proceedings of the Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99), Budapest, Hungary, 5–9 September 1999; pp. 2347–2350.
20. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.
21. Atal, B.S.; Hanauer, S.L. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* **1971**, *50*, 637–655. [CrossRef]
22. Wu, Y.J. *Study on the HMM-Based Speech Synthesis Techniques*; University of Science and Technology of China: Hefei, China, 2006.
23. Cataldo, E.; Leta, F.R.; Lucero, J.; Nicolato, L. Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure. *Mech. Res. Commun.* **2016**, *33*, 250–260. [CrossRef]
24. Schröder, M. Emotional speech synthesis: A review. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 2–7 September 2001.
25. Zen, H.; Nose, T.; Yamagishi, J.; Sako, S.; Masuko, T.; Black, A.W.; Tokuda, K. The HMM-based speech synthesis system (HTS) version 2.0. In Proceedings of the ISCA Workshop on Speech Synthesis, Bonn, Germany, 22–24 August 2007; pp. 294–299.
26. Meng, F.B. *Analysis and Generation of Focus in Continuous Speech*; Tsinghua University: Beijing, China, 2013.
27. Zhuang, X.; Huang, J.; Potamianos, G.; Hasegawa-Johnson, M. Acoustic fall detection using Gaussian mixture models and GMM supervectors. In Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 69–72.
28. Lin, Z.H. *Research on Speech Synthesis Technology Based on Statistical Acoustic Modeling*; University of Science and Technology of China: Hefei, China, 2008.
29. Fan, Y.; Qian, Y.; Xie, F.L.; Soong, F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
30. Zen, H. Acoustic modeling in statistical parametric speech synthesis-from HMM to LSTM-RNN. In Proceedings of the The First, International Workshop on Machine Learning in Spoken Language Processing (MLSLP2015), Aizu, Japan, 19–20 September 2015.
31. Dudley, H. Remaking speech. *J. Acoust. Soc. Am.* **1939**, *11*, 169–177. [CrossRef]

32. Kawahara, H.; Masuda-Katsuse, I.; Cheveigne, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Commun.* **1999**, *27*, 187–207. [CrossRef]
33. HMM/DNN-Based Speech Synthesis System (HTS). Available online: http://hts.sp.nitech.ac.jp/ (accessed on 15 March 2015).
34. Banno, H.; Hata, H.; Morise, M.; Takahashi, T.; Irino, T.; Kawahara, H. Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation. *Acoust. Sci. Technol.* **2007**, *28*, 140–146. [CrossRef]
35. STRAIGHT. Available online: https://github.com/shuaijiang/STRAIGHT (accessed on 25 July 2018).
36. Flanagan, J.L.; Golden, R.M. Phase vocoder. *Bell Syst. Tech. J.* **1966**, *45*, 1493–1509. [CrossRef]
37. McAulay, R.; Quatieri, T.F. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust. Speech Signal Process.* **1986**, *34*, 744–754. [CrossRef]
38. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
39. World. Available online: https://github.com/mmorise/World (accessed on 18 May 2019).
40. Ling, Z.H.; Deng, L.; Yu, D. Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7825–7829.
41. Deng, L.; Seltzer, M.L.; Yu, D.; Acero, A.; Mohamed, A.R.; Hinton, G. Binary coding of speech spectrograms using a deep auto-encoder. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Japan, 26–30 September 2010; pp. 1692–1695.
42. Gehring, J.; Miao, Y.; Metze, F.; Waibel, A. Extracting deep bottleneck features using stacked auto-encoders. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3377–3381.
43. Kang, S.Y.; Qian, X.J.; Meng, H. Multi-distribution deep belief network for speech synthesis. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8012–8016.
44. Zen, H.; Senior, A. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 4–9 May 2014; pp. 3844–4848.
45. Bishop, C. *Mixture Density Networks*; Tech. Rep. NCRG/94/004; Neural Computing Research Group, Aston University: Birmingham, UK, 1994.
46. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM networks. In Proceedings of the IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; pp. 2047–2052.
47. Graves, A.; Fernandez, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
48. Li, R.N.; Wu, Z.Y.; Liu, X.Y.; Meng, H.; Cai, L.H. Multi-task learning of structured output layer bidirectional LSTMs for speech synthesis. In Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; pp. 5510–5514.
49. Yu, Q.J.; Liu, P.; Wu, Z.Y.; Kang, S.Y.; Meng, H.; Cai, L.H. Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages. In Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016; pp. 5545–5549.
50. Caruana, R. *Multitask Learning*; Springer: Berlin, Germany, 1998.
51. Seltzer, M.L.; Droppo, J. Multi-task learning in deep neural networks for improved phoneme recognition. In Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6965–6969.
52. Luong, M.T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. *arXiv* **2015**, arXiv:1508.04025.
53. Jaitly, N.; Le, Q.V.; Vinyals, O.; Sutskever, I.; Sussillo, D.; Bengio, S. An online sequence-to-sequence model using partial conditioning. In Proceedings of the Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 5067–5075.

54. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
55. Wang, W.; Xu, S.; Xu, B. First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention. In Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016; pp. 2243–2247.
56. Bahdanau, D.; Cho, K.; Bengio. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
57. Zhang, J.X.; Ling, Z.H.; Dai, L.R. Forward attention in sequence-to-sequence acoustic modeling for speech synthesis. In Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4789–4793.
58. Oord, A.V.D.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv Preprint* **2017**, arXiv:1609.03499.
59. Oord, A.V.D.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; and Graves, A. Conditional image generation with pixelcnn decoders. In Proceedings of the Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4790–4798.
60. Oord, A.V.D.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel recurrent neural networks. *arXiv* **2016**, arXiv:1601.06759.
61. Oord, A.V.D.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Casagrande, N. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv* **2017**, arXiv:1711.10433.
62. Arik, S.Ö.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Sengupta, S. Deep voice: Real-time neural text-to-speech. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 195–204.
63. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Le, Q. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
64. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Le, Q.V. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv* **2017**, arXiv:1703.10135.
65. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
66. Barron, A. Implementation of Google's Tacotron in TensorFlow. Available online: https://github.com/Kyubyong/tacotron (accessed on 20 October 2018).
67. Ito, K. Tacotron Speech Synthesis Implemented in TensorFlow, with Samples and a Pre-Trained Model. Available online: https://github.com/keithito/tacotron (accessed on 20 October 2018).
68. Yamamoto, R. PyTorch Implementation of Tacotron Speech Synthesis Model. Available online: https://github.com/r9y9/tacotron_pytorch (accessed on 20 October 2018).
69. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4784–4788.
70. Zhang, Y.J.; Pan, S.; He, L.; Ling, Z.H. Learning latent representations for style control and transfer in end-to-end speech synthesis. In Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 6945–6949.
71. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2014**, arXiv:1312.6114.
72. Gibiansky, A.; Arik, S.; Diamos, G.; Miller, J.; Peng, K.; Ping, W.; Zhou, Y. Deep voice 2: Multi-speaker neural text-to-speech. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2962–2970.
73. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Saurous, R.A. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
74. Yasuda, Y.; Wang, X.; Takaki, S.; Yamagishi, J. Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 6905–6909.

75. Nakamura, K.; Hashimoto, K.; Oura, K.; Nankaku, Y.; Tokuda, K. Singing voice synthesis based on convolutional neural networks. *arXiv* **2019**, arXiv:1904.06868.
76. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.O.; Kannan, A.; Narang, S. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In Proceedings of the Sixth International Conference on Learning Representations, Vancouver, BC, Canada, April 30–3 May 2018; pp. 1–16.
77. Chen, H.; Liu, X.; Yin, D.; Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 25–35. [CrossRef]
78. Mametani, K.; Kato, T.; Yamamoto, S. Investigating context features hidden in End-to-End TTS. In Proceedings of IEEE the 44th International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 6920–6924.
79. Chung, Y.A.; Wang, Y.; Hsu, W.N.; Zhang, Y.; Skerry-Ryan, R.J. Semi-supervised training for improving data efficiency in end-to-end speech synthesis. In Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 6940–6944.