Artificial Intelligence Review (2023) 56:5837–5880
https://doi.org/10.1007/s10462-022-10315-0

Conventional and contemporary approaches used in text to speech synthesis: a review

Navdeep Kaur[1,2] **· Parminder Singh**[3]

Published online: 13 November 2022

Abstract
Nowadays speech synthesis or text to speech (TTS), an ability of system to produce human like natural sounding voice from the written text, is gaining popularity in the field of speech processing. For any TTS, intelligibility and naturalness are the two important measures for defining the quality of a synthesized sound which is highly dependent on the prosody modeling using acoustic model of synthesizer. The purpose of this review survey is firstly to study and analyze the various approaches used traditionally (articulatory synthesis, formant synthesis, concatenative speech synthesis and statistical parametric techniques based on hidden Markov model) and recently (statistical parametric based on deep learning approaches) for acoustic modeling with their pros and cons. The approaches based on deep learning to build the acoustic model has significantly contributed to the advancement of TTS as models based on deep learning are capable of modelling the complex context dependencies in the input data. Apart from these, this article also reviews the TTS approaches for generating speech with different voices and emotions to makes the TTS more realistic to use. It also addresses the subjective and objective metrics used to measure the quality of the synthesized voice. Various well known speech synthesis systems based on autoregressive and non-autoregressive models such as Tacotron, Deep Voice, WaveNet, Parallel WaveNet, Parallel Tacotron, FastSpeech by global tech-giant Google, Facebook, Microsoft employed the architecture of deep learning for end-to-end speech waveform generation and attained a remarkable mean opinion score (MOS).

Keywords Concatenative speech synthesis · Formant speech synthesis · Articulatory speech synthesis · Statistical parametric speech synthesis using hidden Markov model and deep learning methods · Expressive TTS · Multi-lingual and multi-speaker TTS · Autoregressive and non-autoregressive models · Speech quality metric · Speech corpus

✉ Navdeep Kaur
nav_783@yahoo.com

Parminder Singh
parminder2u@gmail.com

[1] Department of Computer Science and Engineering, Government Polytechnic College for Girls, Amritsar, India

[2] Research Scholar, Department of Computer Science and Engineering, IK Gujral Punjab Technical University, Kapurthala, India

[3] Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India

1 Introduction

Human beings communicate with each other using different forms of communication where speech is a verbal and most efficient way to communicate (Coelho et al. 2013). Enormous research has been carried out in the speech processing applications that enables computer to understand the written and spoken languages. This extensive research enables computer to speak in the way similar to human and this artificial synthesis of speech waveforms from the written text is known as speech synthesis or TTS (Botha et al. 2012; Lee and Qian 2004). The intelligent TTS has vast assortment of application ranging from commercial to personal use, where commercially it is used for developing network applications, screen readers, multimedia applications, computer vision (Le et al. 2021) and are thus similarly used on personal devices to read documents, learn new languages, creating talking toys and many more (Liang et al. 2004).

The quality of generated waveforms is examined on the basis of its naturalness and intelligibility (Jinyin et al. 2021). Here, intelligibility characterized ease in the production of synthesized speech and its understandability whereas, naturalness refers to its closeness to human speech (Alias et al. 2008). Based on different languages spoken in the world, different TTS exists but TTS built on low resource languages like Punjabi (regional language), which is spoken by approximately 110 million peoples, still lacks in intelligibility and naturalness of the generated speech. Also, with the advancement in deep learning techniques and public availability of large speech corpus (Zhang et al. 2019a, b), the quality of synthetic speech employing corpus for a single speaker is high and close to natural speech (Shen et al. 2018a, b; Saito et al. 2019) but it is difficult to scaled for multi-speakers' speech corpus yet.

Generally, TTS composes of two modules that deals with Natural Language Processing at front end and Digital Signal Processing at back end (Siddhi et al. 2017). The front end of TTS called Text Analysis firstly, reads the written text, performs some preprocessing tasks to covert the written form of text to its verbal form and then generates the linguistic vector for each identified phoneme, syllable of the word as shown in Fig. 1. This front end is highly language specific. The task of back end of TTS (Waveform Synthesis Module) is to artificially synthesize the speech waveforms from the symbolic input linguistic vector (Tabet and Boughazi 2011; Sasirekha and Chandra 2012).

1.1 Motivation

Enormous research has been carried out in the past few decades in the field of speech generation. With progression in technology and evolution in the field of artificial engineering, the deep learning paradigm is successfully applied to generate artificial voice from the

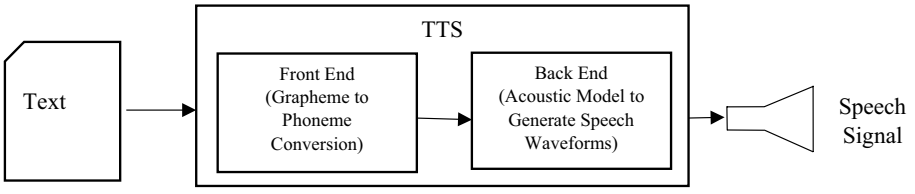


Fig. 1 Block diagram of TTS

written text by a computer. This synthesized voice comprises the characteristics of both intelligibility and naturalness and has been practically used in existent applications. Our motivation in studying and writing this literature review is to examine both conventional and contemporary approaches used to implement speech synthesizer or TTS with their pros and cons. Further, researchers working in the field of speech generation will benefit from this review as it will help them to study various techniques used for TTS.

2 Methodology

A systematic literature review (SLR) involves studying of research papers and review articles relevant for the field of study. An unbiased and well-structured SLR requires to classify and analyze the applicable methodology related to research field before writing a review. The search and selection criteria of papers for writing this review paper are discussed in the following sections.

2.1 Search strategy

Speech synthesis, in a field of speech generation, involves various conventional (formant, articulatory, unit based concatenative speech synthesis, hidden Markov model based statistical parametric speech synthesis) and contemporary (deep learning based) approaches for its implementation. It comprises more than 15 lakh research articles on popular corpora such as Google Scholar when performing search with keyword "speech synthesis." Therefore, the research is confined by first identifying the approaches used for speech synthesis and then by further refining our research on the basis of individual approach mentioned above. In addition, more challenging TTS systems such as emotive, multi-speaker and multi-lingual TTS systems are also included in search.

2.2 Selection strategy

The availability of plenty of research papers and review surveys in diverse areas of research field at various publication venues needs right selection strategy to select papers for writing SLR. Our selection strategy is based on five metrics to ensure that papers are selected from reputed journals only. These are Google Scholar h5 index, SCIMago journal ranking (SJR), journal quartile (Q1–Q4), number of citations of article and year of publication. The distribution of all five metrics from references sources are shown in Figs. 2, 3, 4, 5, and 6.

3 Classification of speech synthesis techniques

Speech synthesis techniques are classified into various categories depending upon the acoustic model used to build the system (Panda et al. 2015a) shown in Fig. 7.

3.1 Articulatory synthesis

The method of Articulatory Speech synthesis synthesizes the sound waveforms based on the natural human speech production system that involves digitally simulating the air flow

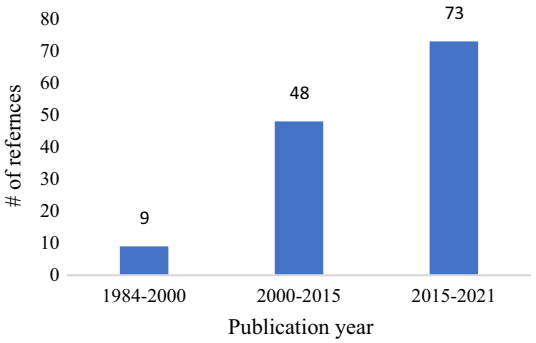


Fig. 2 Distribution of number of references over publication year

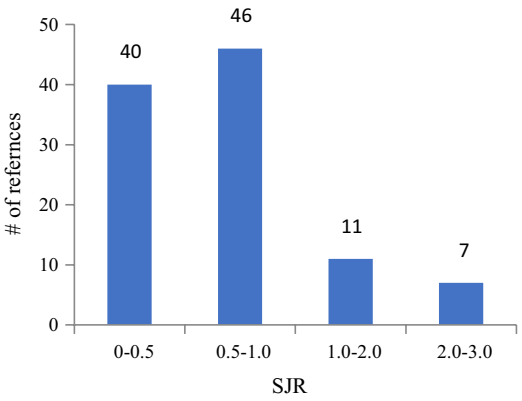


Fig. 3 Distribution of number of references over SJR

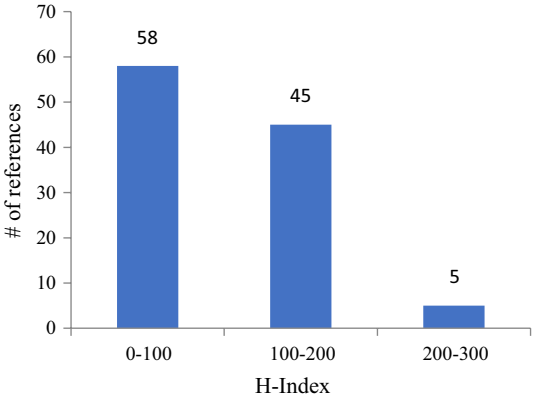


Fig. 4 Distribution of number of references over H-Index

through the vocal tract and the articulation activities happening there (Qinsheng et al. 2011; Peter et al. 2019). Although, this type of synthesis produces highly intelligible speech but it lacks naturalness. The generated waveforms sound robotics. This is due to the reason that data used to model natural human speech production system is derived traditionally from 2-D X-Ray analysis hence, cause difficulty in modeling the articular with precision.

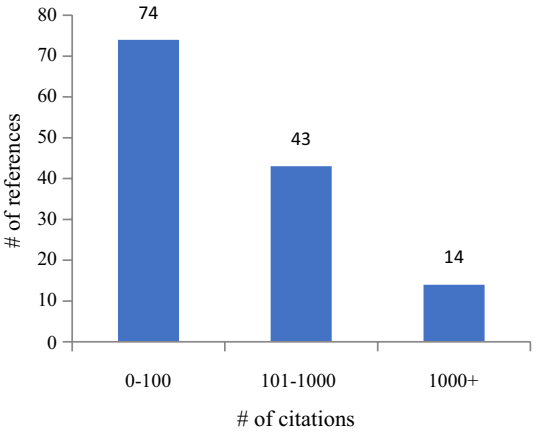


Fig. 5 Distribution of number of references over number of citations

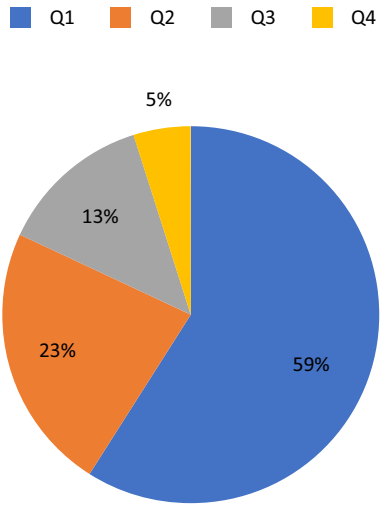


Fig. 6 Distribution of quartile of journals

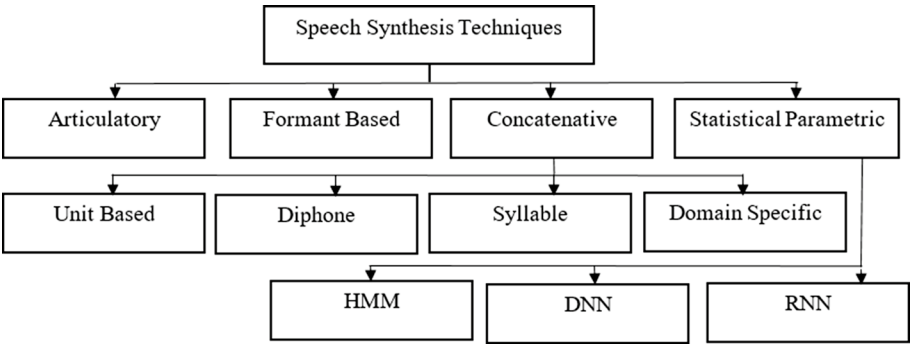


Fig. 7 Various speech synthesis techniques employed to build TTS

Practically, it is also considered as difficult method to implement and therefore, the same level of success was unachievable as compared to other methods of speech synthesis.

3.2 Formant synthesis

Formant speech synthesis is a rule-based approach to synthesize speech used in the past. The acoustic model of formant synthesis used the various sound parameters, like degree of voicing in excitation, fundamental frequency (f0), formant frequencies and their amplitude. Used in both parallel and cascade structures, it produces the highly intelligible speech by avoiding the acoustic glitches at the boundaries of frames/sound units that commonly occurred in unit based concatenative speech synthesis technique but it still generates speech in robotic-sounding form, hence, lacks naturalness of speech (Lukose and Upadhya 2017; Khorinphan et al. 2014).

3.3 Concatenative speech synthesis

TTS systems, based on unit selection, also referred as corpus-based method uses the concatenative speech synthesis approach. In this approach, the synthetic voice is generated by concatenating the different acoustic units together (Gujarathi and Patil 2021). The concatenative speech synthesis is mainly categorized as unit based and diaphone based (Yishuang et al. 2019). The unit selection based concatenative speech synthesis system use many long hours of speech that contains multiple replicas of each diaphone in comparison to diaphone synthesis that uses single copy of each diaphone. Moreover, in diaphone synthesis, the algorithms like pitch synchronous overlap add (PSOLA) are used to change the prosody of concatenated units whereas, in concatenative synthesis, no or negligible signal processing is applied to the speech units (Rao and Narendra 2019). The main problem with diaphone synthesis is that PSOLA algorithm used here for prosody generation requires the pitch period to be labeled precisely on the diaphone units which makes synthesized sound unnatural.

The strength of unit-based concatenative speech synthesis is its massive size speech corpus, which stores almost every utterance spoken in a language on which TTS is built. Each recorded utterance is needed to be fragmented into some discrete phones, syllables, morphemes, words, phrases, and sentences for the purpose of database construction (Kayte et al. 2015). From the available speech corpus, the largest matching sound files are concatenated to generate the target speech as shown in Fig. 8.

The high-quality synthetic voice is generated by using massive database of voice recordings as the sound that is synthesized exists previously in the speech corpus. (Ijima et al. 2015). The objective is to minimize the target and the concatenative cost involved in the searching and concatenating units together while generating high quality speech (Zhou et al. 2021; Karabetsos et al. 2010). The target cost involves matching the candidate unit from speech corpus to the required unit based on prosody and phonetic features whereas, concatenation cost regulates the combination of selected units. The target cost $C^{(t)}$ to select both candidate and required unit is defined as:

$$C^{(t)}\left(t_i, u_i\right) = \sum_{j=1}^{p} w_j^{(t)} C_j^{(t)}\left(t_i, u_i\right), \tag{1}$$

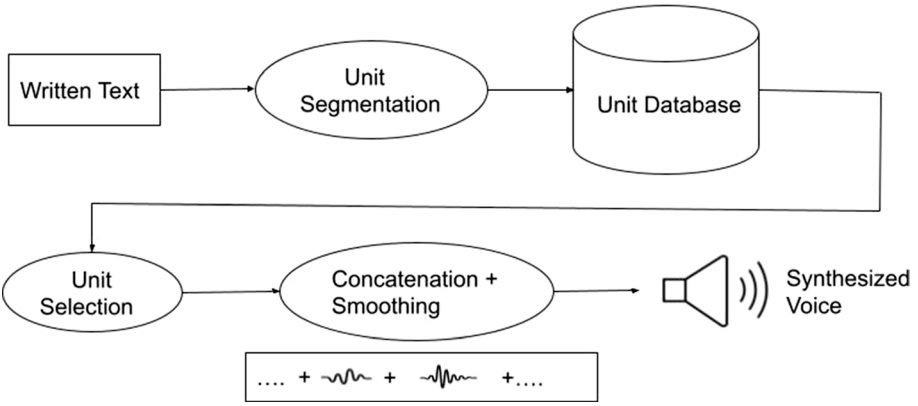


Fig. 8 Concatenative speech synthesis: units are stored in corpus and concatenated to synthesized sound (Kayte et al. 2015)

Here, t_i and u_i represents the candidate unit and the required unit respectively, p is the total count of phonetic and prosodic features undertaken and where $w^t = [w^t_1, w^t_2, \ldots, w^t_P]$ is the relative weight of each sub cost.

The concatenation cost $C^{(c)}$ to concatenate the selected candidate units is defined as:

$$C^{(c)}\left(u_{i-1}, u_i\right) = \sum_{k=1}^{q} w_k^{(c)} C_k^{(c)}\left(u_{i-1}, u_i\right), \tag{2}$$

Both the above-mentioned costs need to be optimized so that optimal sequence of units is determined and overall cost is minimized. Instead of determining the joint cost on one distance measure, it can also be calculated through spectral joint cost (Karabetsos et al. 2010) which involves distance based on the spectral features of speech frame at the concatenation points.

Although high quality speech is regenerated using unit-based speech synthesis but, it requires the huge database of speech units which makes it difficult to implement on limited memory devices like cell phones. To reduce the size of footprint in concatenative speech synthesis, Lee et al. (2007) proposed a novel speaker dependent speech coding and synthesis algorithm. This method utilized a pitch code book created from a corpus of single speaker. Coding efficacy is improved by combing the predictive and non-predictive frame types and the resultant TTS was tested on Korean speech and found 55% lower decoding complexity than G.729 annex A (Salami et al. 1997). Moreover, Sharma et al. (2018) proposed the compressed sensing (CS) and sparse representation (SR) framework to compress the footprint size. As an alternative of storing raw speech waveforms, CS framework employs the storage of signs of CS representations. The proposed CS/SR framework needs to have sparse representation over a learned dictionary to efficiently recreate the speech signals. To determine the effectiveness of new framework, the experiment is conducted using analytical and learned dictionaries that are extracted using K-singular value decomposition, greedy adaptive dictionary (GAD) and principal component analysis algorithms (PCA). Based on voiced and unvoiced type of speech signal, significant coefficients of sparse vector are chosen which, further reduced the size of the footprint. Experimental study performed on Hindi, Rajasthani and on Indian English comprised total of 59, 55 and

43 phonemes respectively. Experimental results showed that natural and intelligible speech was generated using the compressed representations of footprint. The limitation of this method was that it still synthesized reading style voice. For generating voice with diverse speaking styles and emotions, there is a need of especially huge database with different speaking styles voice recordings. It is difficult to collect such a huge data which makes this approach exceptionally time consuming and inefficient (Zen et al. 2007). The synthesized voice is completely dependent on original voice which makes it difficult to change voice characteristics (Takamichi et al. 2014) also. Despite this, the smooth transition at unit boundaries is not obtained.

To take the advantage of both smooth transition among units in statistical TTS (STTS) and natural sound of concatenative TTS (CTTTS), Tiomkin et al. (2011) worked on hybrid text to speech synthesis system (HTTS) by combining both TTS and proposed hybrid dynamic path algorithm in which a cost function is calculated from the spectral distance between consecutive natural units. Experimental results based on listening test showed that sound waveforms generated with HTTS are more natural than by using simple STTS. It also helps in reduction of discontinuities in concatenate speech unit when the size of speech corpus is small.

3.4 Statistical parametric speech synthesis

To control various speech characteristics, the techniques based Statistical Parametric Speech Synthesis (SPSS) approach are employed to build TTS (Zen et al. 2009). In SPSS, the speech parameters like excitation (fundamental frequency) and spectral features are extracted using vocoders from the training database of speech recording and used to generate the speech waveform during synthesis time (Ling et al. 2015).

3.4.1 Hidden Markov model

Tokuda et al. (2013) describes that unit-based speech synthesis produces a speech of high quality by concatenating acoustic units together that are stored in huge database of speech recording. However, the limitation is that it produces reading style speech in the same style without any expression. To generate speech with various styles, there is a need of large database of recordings with different styles which makes this approach very time consuming and inefficient. To change various speech parameters, an approach based on speech parameters called hidden markov model (HMM) based statistical parametric speech synthesis is used to generate speech with different speaking styles. Along with this, HMM based speech synthesis demands less storage requirement than concatenative approach to speech synthesis (Drugman et al. 2009). In 1995, introduced by Tokuda et al. (1995) with the HTS toolkit (HMM-based speech synthesis toolkit), speech synthesis based on HMM model is still dynamic research field (Zen et al. 2007). The various speech parameters are modelled using generative model (Zen et al. 2009). If generative model is HMM then, it is called HMM-based Speech Synthesis. It is based on the principal of formant synthesis and use various speech parameter during training phase and then, involves the generation of speech waveform during synthesis phase (Drugman et al. 2009; Rao and Narendra 2019; Yishuang et al. 2019). HMMs and decision trees (Yoshimura et al. 1999; Narendra and Rao 2017) are the two main components of this method.

3.4.1.1 Decision trees and H(S)MMs HSMM (Hidden Semi-Markov Model) a variant of HMM is used to model speech excitation and the spectrum information. Each phoneme of text in HTK toolkit is modeled by one HMM and complete sentence is synthesized by concatenating various HMMs comprises the sentence. The two non-emitting states at the beginning and end of each HMM are used for concatenating HMMs together. Decision trees are used to model prosodic and linguistic features of speech and individual node of decision tree is tied to one contextual feature.

3.4.1.2 Synthesis process The speech excitation feature such as f0 and spectral features such as Mel frequency cepstral coefficients (MFCC) which are extracted from the speech dataset are used to train context dependent HMM as described in Fig. 9. In speech processing applications, MFCCs and spectrogram are the prevalent choice of output features (Latif et al. 2020). During synthesis, feature vector for acoustic observation $O = [0_1{}^{\mathrm{T}}, O_2{}^{\mathrm{T}}, \ldots, O_{\mathrm{T}}{}^{\mathrm{T}}]^{\mathrm{T}}$ is created from the speech parameters, where $(\cdot)^{\mathrm{T}}$ signifies matrix transpose and T refers to the total frames used. (Tokuda et al. 2000). In HTS System, for example, this vector comprises: MFCC coefficients, f0, aperiodicity coefficients (Zen et al. 2007), and prosodic parameters. Lastly, to generate speech waveforms, the speech parameter trajectories are passed to a vocoder, usually STRAIGHT, a high-quality vocoder (Kawahara et. al.

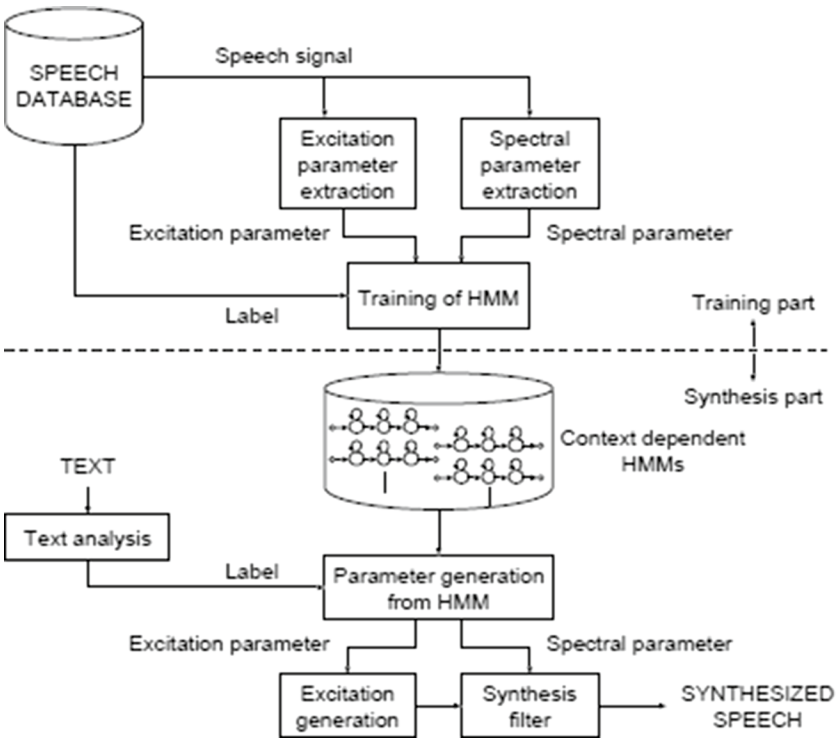


Fig. 9 A block diagram showing the training and synthesis processes in HTS—HMM-based speech synthesis system (Zen et al. 2007)

1999) or speech signal processing toolkit (SPTK) (Fukada et. al. 1992) to synthesize sound waveforms.

Since quantity of training data is not adequate to model each combination of context dependent HMMs now, the top-down decision trees with clustering are being used. In this clustering approach, the clusters are formed by grouping the states of context dependent HMMs. The process of assigning the HMMs to each cluster is done by inspecting the context combination of individual HMM using a binary decision tree. These decision trees link context dependent binary question to their non-terminating node. The minimum description length (MDL) criterion is then used to fix the length of binary decision trees (Shinoda and Watanabe, 2000). The advantage of using the HMM based synthesis over unit-based synthesis is that a synthetic speech with varying speaking styles can be generated by using smaller footprints (Ling et al. 2006; Zen et al. 2007).

Although it has an advantage over concatenative speech synthesis as only small corpus is needed to learn a voice, yet to have voice that resembles human voice it needs few hours of speech in its training corpus. Adequate learning data in speech corpus results in better acoustic modeling which further results in synthesizing natural sounding speech at the synthesis time. The speech corpus can be from multiple speakers (Yamagishi et al. 2009) in contrast to single speaker database used in concatenative method. The main limitations of HMM based synthesis are the inaccuracy of acoustic modelling that is statistical averaging process used in parametric methods generate smooth speech trajectories which lead to muffled speech (Zen et al. 2009; Tokuda et al. 2013), and wrongly extraction of pitch information (Reddy and Rao 2017). To improve the acoustic modeling so that effect of over-smoothen speech trajectories can be reduced, several techniques have been proposed in literature like global variance (GV) based parameter generation (Toda and Tokudu 2007; Toda 2011), modulation spectrum-based post filtering (Takamichi et al. 2014), modified discrete cosine transformation (MDCT) (Biagetti et al. 2018) and enhancement in spectral tilt of generated speech (Sharma and Parsanna 2017). These approaches are capable to enhance the intelligibility of speech primarily at segment of spectral level only (Yin et al. 2015). Although the well-known pitch extraction algorithms (Reddy and Rao 2017; Kawahra et al. 1999; Talkin 1995) can precisely determine the pitch from the speech signal at the area of high amplitude but for a creaky sound with a low amplitude area, these algorithms erroneously detect the voiced and unvoiced regions. In addition, the decision tree-based context dependent HMMs generates the speech of reasonable quality but it suffers from many limitations. Decision trees are not suitable to model the complex context dependencies. They will grow large and fragment the training data which degrades the quality of the generated speech (Zen et al. 2013). Moreover, as too many contextual factors for a language are under consideration, their combination increases exponentially (Tokuda et al. 2002).

3.4.2 Speech synthesis based on deep learning methods

Zen et al. (2013) mentioned the shortcomings of HMM systems and proposed the deep learning techniques. To tackle the issues of HMM based synthesis related to decision trees and to improve the quality of synthesized voice, new methods based on deep neural networks (DNN) (Qian et al. 2014; Zen et al. 2013) or recurrent neural networks (RNN) (Achanta et al. 2017) have been used in recent years for the purpose of learning. The speech synthesizer based on DNN are also used in various contemporary techniques like voice conversion (Sisman et al. 2021; Huang et al. 2021; Chen et al. 2014b), emotion

synthesis and regeneration of dialects (Tits et al. 2019; Zhang et al. 2019a, b). Further, a variant of DNN known as long short-term memory recurrent neural networks (LSTM-RNNs) showed the best results in the field of speech processing (Zen and Sak 2015). The practice of using neural networks in the area of speech processing such as ASR and TTS is not novel, and plenty of work has been done in the era of 80–90s to model speech and its components. The process is further expediting with the remarkable progress in the field of hardware and software that enable the DNN to train with a substantial amount of training data. Thus, attained a significant progress over the conventional approaches to speech processing including the areas of speech recognition (Ghajabi and Herando 2018), machine translation, singing voice synthesis, speech synthesis and many others (Hinton et al. 2012; Sutskever et al. 2014; Nakamura et al. 2019; Mametani et al. 2019). However, the recent advancement both in hardware (e.g., graphical processing unit (GPU)) and software empowers the researchers to train a DNN from a significantly vast quantity of training data.

4 Elementary models used in deep learning

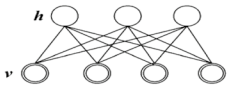
Since last two decades, techniques based on deep learning are used in numerous fields especially, related to signal processing such as speech synthesis, speech recognition, spectrogram coding, dialog management systems and many more (Ling et al. 2013). The following section gives the review of elementary models used for deep learning.

4.1 Restricted Boltzmann machines

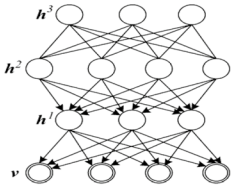
Mainly used as pre-trainer for DNNs (Zoughi and Homayoonpoor 2018), the restricted Boltzmann machines (RBM) is a two-layered graphical model that is used to map dependency among random variables (Nakashika et al. 2019; Chin et at. 2013) without supervision (Nakashika and Yatabe 2021). These two layers are known as visible and hidden layers of network with intra-connection among the nodes of same layer shown in Fig. 4. By using different energy functions, the RBMs are capable of modeling real valued data often used in speech synthesis and speech recognition tasks, categorical data, binary data and mixed of real, binary and categorical data (Ling 2015; Dong et al. 2021). For the visible stochastic layer with nodes $V=[v_1, v_2,\ldots,v_n]^T$ and hidden stochastic layer with units $H==[h_1,h_2,\ldots\ldots h_m]^T$ where, n and m signifies the counts of units at each layer respectively, the energy functions E when units are binary is stated as:

$$E(v,h:\lambda)=-\sum_{i=1}^{V}a_i v_{i-}\sum_{j=1}^{H}b_j h_j-\sum_{i=1}^{V}\sum_{j=1}^{H}w_{ij}v_i h_j \tag{3}$$

Here, weights between visible unit (v_i) and hidden unit (h_i) are represented using matrix w_{ij}, a and b denotes their biases respectively and λ denotes the set of model parameter comprises a, b and w_{ij}. For the purpose of speech synthesis, the maximum likelihood estimation (MLE) is used to determine parameters of RBM and then, spectral envelope modeling is accomplished using RBM-DNN network structure. As speech parameters are real-valued, various authors Ling et al. (2013), Chen et al. (2015) and Hu et al. (2016) all work on RBMs to improve the intelligibility of synthetic voice based on deep learning methods as



(a) RBM with visible layer v and hidden layer (Ling et al. 2013)



(b) DBN with stacked RBM (Ling et al. 2013)

Fig. 10 Deep Belief Network Structure. **a** RBM with visible layer v and hidden layer (Ling et al. 2013). **b** DBN with stacked RBM (Ling et al. 2013)

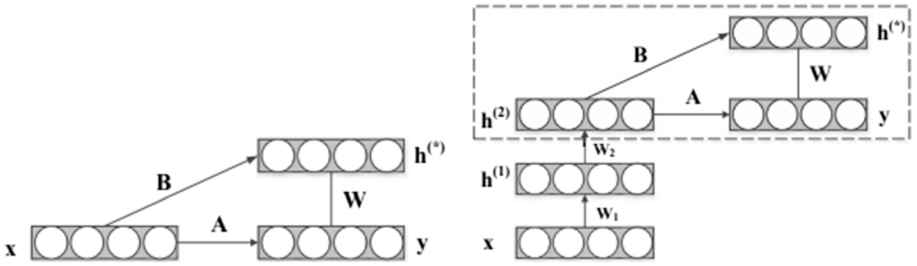


Fig. 11 CRBM and two-hidden layer DCRBM (Yin 2016)

RBM works efficiently when $v \in R^V$ are real-valued and $h \in \{0,1\}^H$ are binary. Such form of RBM is known as Gaussian-Bernoulli form (Ling et al 2013, 2015).

4.2 Deep belief networks

Being a class of deep neural network, deep belief networks (DBN) are formed by stacking multiple RBMs together as in Fig. 10. Thus, DBN is a network that comprises many hidden units where, units at consecutive layers are linked together and there is no connection among units of own layer. Figure 11 below shows graphical representation of the RBM and three-layer DBN structure where, two layers at the top form the undirected graph with directive to bottom layers to regenerate the visible unit at visible layer. Given a set of visible units and hidden units, the joint probability function is given as:

$$P\left(v, h^1, h^2, \ldots, h^L\right) = P\left(v|h^1\right)P\left(h^1|h^2\right)\ldots P\left(h^{L-2}|h^{L-1}\right)P\left(h^{L-1}|h^L\right) \tag{4}$$

Here, $h^l = [h_1^l, \ldots ..h_{H_l}^l]^T$ denotes hidden stochastic vector of the lth hidden layer, H_l is the dimensionality of h^l and L represents the total hidden layers.

To mitigate the effect of over-smoothing speech trajectories in HMM modeling, Ling et al. (2013) tried to improve the acoustic model by modeling parameters of spectral envelope using RBM-DBN model structure replacing the Gaussian mixture models. These parameters are determined using well-known STRAIGHT vocoder. For the purpose of conducting experiment, 50 units at each layer were taken and each RBM was trained layer-by-layer manner following greedy learning algorithm. The results showed the significant improvement in naturalness of sound and helped to reduce the over-smoothing effect with the average spectral distortion of 4.10 in db. Similarly, Chen et al. (2015), proposed a DNN

based probabilistic postfilter to cope with the muffled sound. The two postfilters worked in Mel-cepstral domain and high dimensional spectral domain are trained using two cascading RBMs with bidirectional associative memory (BAM). The synthetic voices built on male and female speakers showed the remarkable improvement in segmental quality of speech. Further, for a linguistic feature to which speech is needed to be synthesized, the new model that used the output of DNN conditioned on RBM is used to model the probability distribution of spectral envelopes (Yin 2016) and such model is named as deep conditional restricted Boltzmann machine (DCRBM) and shown in Fig. 11.

Here, x is a conditional vector and y and h forms the visible and hidden layers of RBM respectively. A, B and W are the set of weights among the layers. As shown in Fig. 11 the links between layers x and y, and, x and h are directed whereas links between layers forming the RBM is undirected. Given a X, H and Y, the total number of units at each layer with $h = \in \{0,1\}^{H}$, $x \in R^{X}$, and $y \in R^{Y}$, the energy function of CRBM is defined as:

$$E(y,h,x;\theta_c) = \sum_{i=1}^{D_Y} \frac{\left(y_{i-}a_i - \sum_k A_{ki}x_k\right)^2}{2} - \sum_{j=1}^{H}\left(b_j + \sum_k B_{kj}x_k\right)h_j - \sum_{i=1}^{D_Y}\sum_{j=1}^{H} w_{ij}h_j y_i \quad (5)$$

Here $\theta_c = \{W, A, B, a, b\}$ are the parameters of DCRBM. a and b are the biases of both visible and hidden layers respectively. The resultant model is applied on Chinese speech corpus results in better quality of speech with sharp formants and lessened over-smoothing effect having average structure distortion of 3.74 in db. The DCRBMs are proposed to model time-based reliance among the human motion features (Xie et al. 2018) and then later used in the field of voice conversion to model the input and output speech parameters (Wu et al. 2013).

Figure 12 below shows the comparison of mean opinion score (MOS) rating of baseline HMM, simple DBN and using mix approach that combines which use DBN for mel-generalized cepstral coefficients (MGC) and HMM for log f0 using the Multi-distribution DBN (Kang et al. 2013). The higher MOS is found for mix structure as less distorted spectrum is generated. The synthesized voice is more cleared due to lively prosody and smooth f0 pattern in this approach.

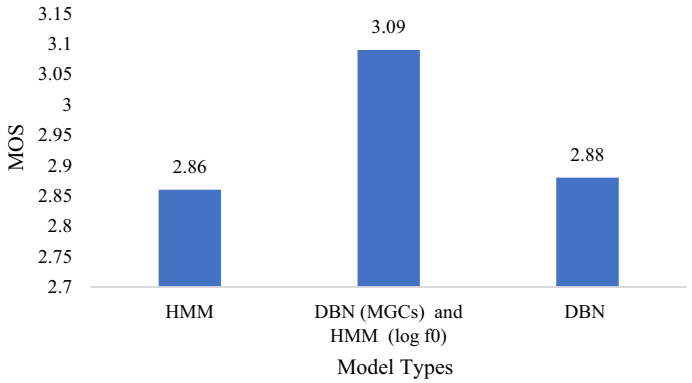


Fig. 12 Comparison of HMM with DBN structure

4.3 Deep neural network

Being a class of artificial neural network, deep neural network (DNN) is a feed-forward neural network which has many hidden layers between the input and output layers (Hinton et al. 2012; Norvig and Russel 2020; Goodfellow et al. 2016; Schmidhuber 2014). Values at each layer are propagated forward to next higher layer using weighted sum of the inputs it receives using a non-linear activation function. DNNs are used for both solving regression and classification problems. As output at each layer is activated using non-linear function, they are capable of modeling complex relationships between input and outputs. However, as the error signal is propagated downward to the network, it causes problems in training a large network composed of many hidden layers as gradient starts vanishing thereby limiting the information available to the lower layers (Kolen and Kermer 2001).

4.3.1 Basic principle of working of DNN based speech synthesis

In speech synthesis based on DNN, the decision trees used in HMMs are replaced by deep neural networks which map the input linguistic feature vector to the output speech parameters to synthesized resultant speech (Nazir and Malik 2021). Excitation feature fundamental frequency (f0) and spectral features are extracted from the speech databases during training phase (Yin et al. 2016). By setting the predicted output features from the DNN as mean vectors and pre-computing global variances of output features from all training data, the speech parameter generation algorithm (To 2007) can generate smooth trajectories of speech parameter features which conciliate both the static and dynamic features of speech. Figure 13 (Zen 2013) showed the basic architecture of using DNN to synthesize voice. In this figure, the text to be articulated first translated into input linguistic vector. This vector is mapped to output speech parameters using deep neural network and lastly, a vocoder at waveform synthesis module synthesized the speech waveform.

The preference test score results of HMM and DNN are compared in Table 1 below which showed the statistical significance of DNN over HMM.

Based upon the representation of input and output speech feature vector, the deep learning methods are classified as deep generative models (cluster to feature mapping), deep joint model and deep conditional model (input to feature mapping) (Ling 2015). DNN is trained using supervised learning algorithms like gradient descent back propagation algorithm. But such algorithms suffer from the problem of overfitting the network (Wang et al. 2021). Despite this, objective measures of experimental results preferred HMM over DNN system. The reason is the interpolation has been done to generate missing f0 values for unvoiced frames which cause noise during DNN training (Yin et al. 2016). To overcome the problem of overfitting, the pre-trained DNN using DBNs have also been proposed in literature (Hinton et al. 2006). To improve the accuracy of acoustic modelling of DNN based systems, restricted Boltzmann machines (RBM) are used to model the spectral envelopes (Yin et al. 2016). Instead of initializing the weights of DNN randomly, the DNN is pertained using the deep belief networks with stacked RBM (Qian 2014). The subjective measures of experimental results preferred the DNN system over HMM system. Another advantage of using DNN over HMM is that they are capable of representing dependencies among the input contextual and output acoustic features of speech automatically (Koriyama and Kobayashi 2019). Moreover, DNN can be easily trained using mini-batch optimization from the large dataset. Nevertheless, the performance of deep neural networks on multiple speaker speech corpus is enhanced by feeding speaker codes in the hidden layers

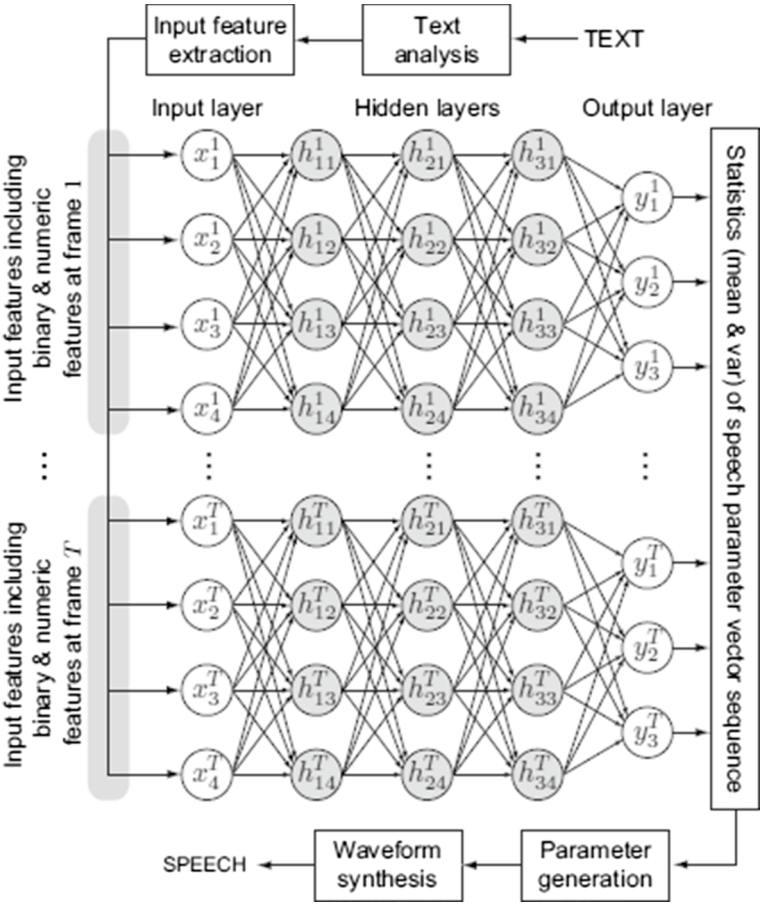


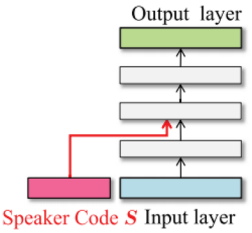
Fig. 13 DNN based Speech Synthesis System (Zen 2013)

Table 1 Preference score test results on HMM and DNN (Zen 2013)

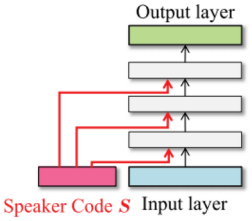
HMM	DNN (Layers × units)	Neutral	p value	z score
15.8	38.5 (4×256)	45.7	$<10^{-6}$	− 9.9
16.1	27.2 (4×512)	56.8	$<10^{-6}$	− 5.1
12.7	63.6 (4×1024)	50.7	$<10^{-6}$	− 11.5

of DNN network structure (Hojo et al. 2018). As shown in Fig. 14, speaker codes for each target speaker are supplied to network to synthesize speakers voice during the process of waveform synthesis. These speaker codes are fed to either a single hidden layer or multiple hidden layers.

The speaker code (ID) vector $v = [v_1, v_{2,\ldots,} v_k]^T$ for speaker m is determined using fixed 1-of-K form:



(a) One-hot speaker code connected to a single hidden layer (Hojo et al. 2018)



(b) One hot speaker code connected to all hidden layers (Hojo et al. 2018)

Fig. 14 Different Methods to connect hot speaker code to hidden layers. **a** One-hot speaker code connected to a single hidden layer (Hojo et al. 2018). **b** One hot speaker code connected to all hidden layers (Hojo et al. 2018)

$$v_k = \begin{cases} 1 & (k = m) \\ 0 & (k \neq m) \end{cases} \tag{6}$$

Here, K is the total speakers in the training data. This model structure generates more natural speech subject to condition of using small number of target speaker utterances.

4.4 Recurrent neural networks

Due to inherit restraint of DNN, these are unable to capture the temporal features thus lead to discontinuities in the predicted speech parameter frame-wise. Although such discontinuities are alleviated using postfiltering techniques in which predicted parameters are optimized through minimum generation algorithms (MGE) but, results in irregularities in training and testing data (Achanta et al. 2017). Alternative to DNNs, RNNs are now recently used for the speech processing applications. In RNN, units are connected to form a cycle. Having a long short-term memory (LSTM) in RNN, they are capable of capturing information from the feature sequence therefore, can map the linguistic text feature vector into the speech output features. It comprises input gate to determine when to recall the input, output gate to send the output and forget gate to determine whether to forget or recall the values (Liu et al. 2018). As described in Fig. 15 (Wen et al. 2018) the left side of figure represented the typical DNN model which comprises few non-linear hidden layers constituted by pre-trained RBMs and one output layer. The right side of figure showed the stacking of at least one recurrent layer comprises bidirectional RNN (BRNN) with LSTM.

It is initially proposed to resolve the problem of vanishing gradient which often arises in DNN. RNNs are used to improve the acoustic model by using categorical and numerical linguistic features for input. Moreover, in comparison with DNN, RNN with bi-directional LSTM (BLSTM) can easily learn the time-series data hence, capable of modelling the co-articulation features in statistical speech synthesis. As a result, these generate smooth speech trajectories at output layer which then synthesize high quality speech (Kolen and Kermer 2001). The research by authors (Wen et al. 2018) show that the BLSTM—RNN outperforms DNN based speech synthesis particularly, the root mean square value (RMS) of log fundamental frequency (LF0) is minimized by 4.8%. For the conduct of experiment, authors implement the two BLSTM-RNN models where the former is employed with multitask learning to predict line spectrum pair (LSP) and unvoiced/voiced (U/V) decision from the contextual features of input vector whereas the latter is used to predict LF0 of the

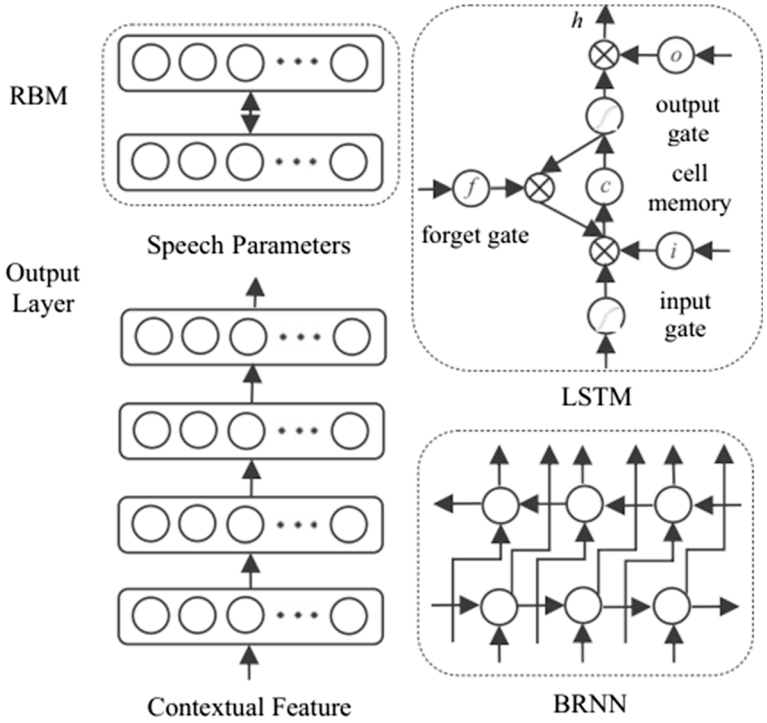


Fig. 15 RNN structure to model speech parameters (Wen et al. 2018)

voiced phonemes as LF0 values only exist for the voiced frames. They also concluded that due to heavy computation involved in the implementation of BLSTM-RNN, this is not easily applicable to the real time speech synthesis applications which is the main drawback of BLSTM-RNN over DNN based speech synthesis (Wang et al. 2018a).

4.5 Deep Gaussian process

DNN models primarily lay emphasis on fitting of data during training phase which though may significantly improve the quality of produced speech but the network eventually suffers from the problem of overfitting (Mitsui et al. 2020). Deep gaussian process (DGP) is a substitute to DNN or RNN, which comprises cascaded bayesian kernel regressions (BKRs) to map the non-linear transformation between input linguistic vector and the output acoustic variables using less numbers of hyperparameters in contrast to large sum of parameters in DNNs (Yang and Klabjan 2021). While training both the model complexity and data fitting is taken into consideration and maximizing marginal likelihood is followed in the process of data fitting which makes the model less prone to overfitting (Koriyama and Kobayashi 2019; Chai et al. 2019). The TTS based on DGP not only outperforms for a single speaker modeling but also works well on multi-speaker modeling (Mitsui et al. 2020, 2021). When introduced, they worked well on small amount of training data (Damianou and Lawrence 2013). But recent research has also shown that DGPs also worked fine on large amount of training data (Cutajar et al. 2017). TTS based on DGPs produces more

natural sound when compared to feed-forward DNNs (Moungsri et al. 2018). Although DGPs do not have recurrent structures still they outperformed LSTM-RNN, and performance of TTS based on DGPs further improved by introducing recurrent architecture. But for this, computational complexity should be a main concern. In DGP, the modeled relationship between input x and output y by GPRs is stated as:

$$\mathrm{y} = \mathrm{f(x)} + \in \tag{7}$$

$$\mathrm{f} \sim \mathrm{GP(m(x), k}\left(\mathrm{x}, x'\right) \tag{8}$$

Here, $\in, m(x)$ and $k\left(x, x'\right)$ are the random noise, mean, and kernel functions respectively. For multi-dimensional output, multiple GPRs are considered.

Radial basis function (RBF) kernel: Being a stationary kernel, RBF is stated as:

$$k_{RBF}\left(x, x'\right) = \exp\left(-\frac{r^2}{2}\right) \tag{9}$$

where,

$$r = \sqrt{(x, x')^T \Lambda^{-1} (x, x')} \text{ and } \wedge = \mathrm{diag}\left[l_1^2, \ldots, l_D^2\right] \tag{10}$$

The output obtained from the RBF kernel is calculated from the distance among the input vectors. RBF kernel suffers from vanishing gradient problem as it approaches zero with increase in distance between input vectors.

Rational quadratic (RQ) Kernel: RQ kernel is stated as the summation of infinite RBF kernels and is represented by:

$$k_{RQ}\left(x, x'\right) = \left(1 + \frac{r^2}{2\alpha}\right)^{-\alpha} \tag{11}$$

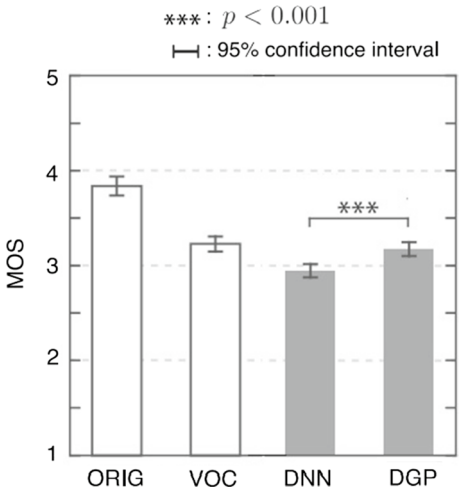


Fig. 16 Comparison of DGP with DNN and VOC

Table 2 Subjective preference test score on various models

VOC	DNN	DGP	LSTM-RNN	p value	Z score
63.3	36.7			$<10^{-4}$	5.18
54.7		45.3		$<10^{-4}$	4.79
	35.7	64.3		0.104	1.62
		55.7	44.3	$<10^{-4}$	5.18

where, $\alpha>0$ represents contour of RQ kernel. The gentle contour is formed when α takes small value and it resembles to RBF kernel when $\alpha \rightarrow \infty$.

When compared to DNN and LSTM-RNN speech synthesis models, DGP is considered to be better in both MOS rating and preference test score (Koriyama and Kobayashi 2019). Figure 16 depicts the comparison of MOS rating of DGP with DNN and vocoded speech samples (VOC). The results also show that there is less difference between VOC and DGP as compared to DNN and DGP. The results of subjective preference test on these models are shown in Table 2 which proved the statistical significance of DGP-based framework over other networks. Figure 17 shows that the distortion is found in various acoustic features in DNN and DGP models.

4.6 Comparison between various methods of speech synthesis

Table 3 below shows the comparison among different types of approaches used for the purpose of speech synthesis along with their advantages and shortcomings.

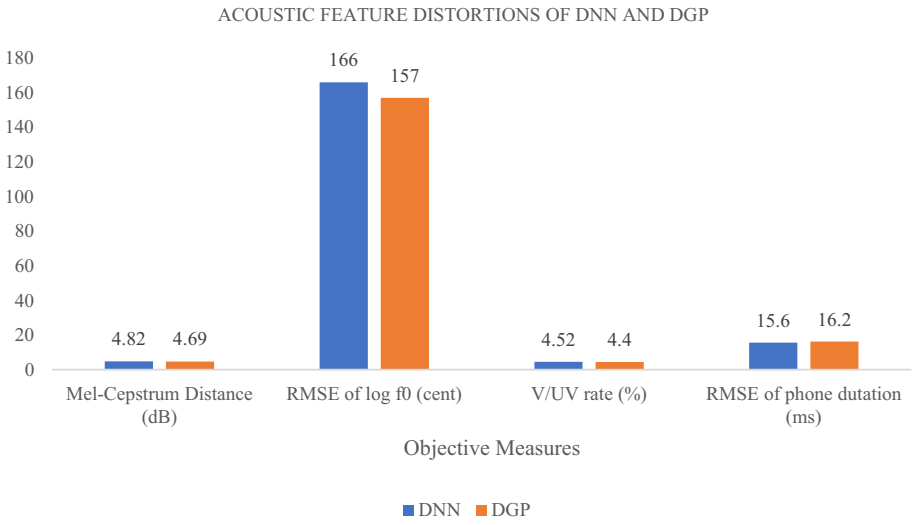


Fig. 17 Acoustic feature distortion of DNN and DGP

Table 3 Comparison of various methods used for speech synthesis

Method	Approach followed	Advantages	Disadvantages
Articulatory synthesis	Mimics the natural speech production system of humans	Intelligible speech is generated; No speech database is needed. (Qinsheng et al. 2011)	Generated speech sounds a robotic as data for articulatory synthesis derived from 2D X-ray images In addition to this, it gains little popularity due to its difficult implementation (Peter et al. 2019)
Formant synthesis	Rule based approach to speech synthesis that generates sound from the stated rules	Generates intelligible speech; prevent acoustic glitches at frame boundaries that is commonly found in concatenative speech synthesis; appropriate for system with limited power and storage; No speech database is needed. (Lukose and Upadhya 2017)	Although highly intelligible speech is synthesized but naturalness of speech is low; Generated speech sounds are artificial and robotic. (Lukose and Upadhya 2017)
Concatenative synthesis	Acoustic units are joined together to generate sound. Use words, syllables, half-syllables, phonemes, di-phones or tri-phones as an acoustic unit. (Gujarathi and Patil 2021; Kayte et al. 2015)	Synthesized speech is close to human sound (Zhou et al. 2021)	Requires large database of acoustic units (Zen et al. 2007); Voice characteristics can't be changed; Reading style speech is synthesized (Takamichi et al. 2014)' Quality of artificial sounds degrades at the concatenation points
HMM based	Statistical parametric based approach. Use HMM as a generative model to generate speech parameters and then, vocoder to generate sound from the generated parameters (Zen et al. 2009; Drugman et al, 2009; Rao and Narendra 2019; Yishuang et al. 2019)	As speech is synthesized from the speech parameters therefore, voice features can be easily modified. Used to generate speech with different speaking style such as emotional speech can be synthesized from lesser quantity of training data (Drugman et al. 2009; Tokuda et al. 2013; Yamagishi et al. 2009))	Suffer from over-smoothing of speech trajectories which leads to muffled speech (Zen et al. 2009; Tokuda et al. 2013). Large training data is needed to generates natural sounding speech. Complex context dependencies are not modeled as decision trees used in HMM suffers from fragmentation problems (Zen et al. 2013; Tokuda et al. 2002)

Table 3 (continued)

Method	Approach followed	Advantages	Disadvantages
HMM using RBM and DBN	RBM and DBN with many hidden layers are used to model the distribution of spectral envelope of speech individually at HMM state During the process of synthesis, spectral envelope is predicted using RBM-HMMs or DBN-HMMs by using maximum probability parameter generation (MPPG) (Ling et al. 2013; Chen et al. 2015; Hu et al. 2016)	Use of RBM-DBN to model low-level spectral envelopes alleviate the problem of over-smoothing from the synthesized speech (Ling et al. 2013; Chen et al. 2015; Hu et al. 2016)	Although introduction of more hidden layer in RBM-DBN advances the accuracy of model but naturalness of sound does not improve accordingly
DNN	Decision trees used in HMM model is replaced by deep neural networks. DNN is used to map the input linguistic variables to the output speech parameters (Zen 2013)	It can both model the complex context dependencies among input features and do not suffer from the fragmentation problem of training data (Zen 2013)	Back propagation algorithm used to train model requires complex matrix multiplication computation at each layer. Also, weights of DNN are not easy to infer (Mitsui et al. 2020). DNNs are not able to model multi-modal property by which computer can speak different texts in different ways (Hojo et al. 2018; Wang et al. 2021)
RNN	A bi-directional long short-term memory is used along with the neural structure to model the speech parameters (Wen et al. 2018; Liu et al. 2018)	Improves the acoustic modeling by capturing the temporal features. (Achanta et al. 2017; Wen et al.2018; Liu et al. 2018)	High computational complexity makes it difficult to implement on real time speech synthesis (Wen et al. 2018)
DGP	Use cascaded Bayesian Kernel Regressions (BKRs) to map the non-linear transformation between input linguistic vector to the output acoustic variables (Yang and Klabjan 2021)	Use a lesser number of hyper-parameters as compared to DNN. Reduces the over-fitting problem of the model. Work well on multi-speaker speech corpus (Mitsui et al. 2020, 2021)	Due to absence of recurrent structures, DGP cannot capture the temporal features Distortion of Mel-cepstrum, f0 and V/UV rate is still higher than DNN

4.7 Metrics used to evaluate performance of TTS

For progression in development of speech processing systems, a "good" quality measure is a key. A speech quality can be measured both subjectively which involve human perception to speech and objectively which involve mathematical measures to speech quality.

4.7.1 Objective test metrics

4.7.1.1 Itakura-Saito measure If $s(i)$ and $s'(i)$ are two frames of speech then $x_n(i)$ and $x'_n(i)$ are the two windowed frames obtained by applying a window function $w(i)$ to speech signal at instance n.

$$x_n(i) = w(i)s(i+1) \tag{12}$$

$$x'_n(i) = w(i)s'(i+1) \tag{13}$$

If $X_n\left(e^{jw}\right)$ and $X'_n\left(e^{jw}\right)$ are the signal obtained after applying the Fourier transform $\ddagger = e^{jw}$ on windowed frames then for each pair of $X_n\left(e^{jw}\right)$ and $X'_n\left(e^{jw}\right)$, spectral distortion $\rho\left[X_n, X'_n\right]$ is stated as dissimilarity among $X_n\left(e^{jw}\right)$ and $X'_n\left(e^{jw}\right)$. For analysis of speech Itakura-Saito measure (Juang 1984) is defined as:

$$\rho_{\mathrm{IS}}\left[X_n, X'_n\right] \triangleq \int_{-\pi}^{\pi} \left[e^{\bigwedge(w)} - \bigwedge(w) - 1\right] \frac{dw}{d\pi}, \tag{14}$$

where,

$$\bigwedge(w) = log\left|X_n\left(e^{jw}\right)\right|^2 - log\left|X'_n\left(e^{jw}\right)\right|^2 \tag{15}$$

4.7.1.2 Root mean square (RMSE) of log f0 Root mean square error (Wang et al. 2008) is used to compare log f0 trajectories generated by TTS system. Mathematically, it is defined as:

$$\sqrt{\sum_{i=0}^{n} \left(\log f_0(i) - \log f_e(i)\right)^2 \Big/ 2} \tag{16}$$

where n is the total number of frames that comprises a sentence, $\log f_0$ is the original and $\log f_e$ are the synthesized f0 contour. Only voiced frames are used for the calculation of RMSE.

4.7.1.3 Gross pith error (GPE) GPE (Babacan et al. 2013) is the measure of proportion of frames which are measured as voiced for both natural and synthesized speech having relative pitch value greater than set threshold value (generally taken as 20% in speech analysis).

4.7.1.4 Voiced/unvoiced (V/UV) error rate It is the measure of the proportion of frames for which an incorrect voiced/unvoiced decision is made with total number of frames. It is measured in percentage.

4.7.1.5 Correlation coefficient It measures how strongly synthesized log f0 is related to natural log f0. The commonly used Pearson correlation coefficient r is defined as:

$$r = \frac{n\left(\sum_{i=1}^{n}(\log f_{0(i)} \log f_{e(i)}\right) - \left(\sum_{i-1}^{n} \log f_{0(i)}\right)\left(\sum_{i-1}^{n} \log f_{e(i)}\right)}{\sqrt{\left[n\sum_{i=1}^{n} \log f_{0(i)}^2 - \left(\sum_{i=1}^{n} \log f_{0(i)}\right)^2\right]}\sqrt{\left[n\sum_{i=1}^{n} \log f_{e(i)}^2 - \left(\sum_{i=1}^{n} \log f_{e(i)}\right)^2\right]}} \quad (17)$$

where n is the total frames, $\log f_0$ and $\log f_e$ are the natural and synthesized f0 contour. if $r = +1$, strong positive relation, if $r = -1$, strong negative relation, if $r = 0$, no relation.

4.7.1.6 Mel cepstral distortion (MCD) MCD measure is used to measure the closeness of synthesized mel cepstra sequences with the natural sound. MCD is computed either by aligning the two sequences in terms of their timing or using dynamic wrap timing (DWT). The mean MCD of two waveforms synthesized w^s and ground truth w^{gt} is measured as (Kominek et al. 2008):

$$\mathrm{MCD}\left(w^s, w^{gt}\right) = \frac{\alpha}{T'}\sum_{t=0}^{T-1}\sqrt{\sum_{d=s}^{D}\left(w_d^s(t) - w_d^{gt}(t)\right)^2} \quad (18)$$

such that $ph(t) \notin \mathrm{SIL}$.

where $\alpha = \frac{10\sqrt{2}}{\ln 10} = 6.14185$, d is dimensional index ranging from 0 to 24, t is the time or frame index, t is given as $\min(|w^s|, |w^{gt}|)$ such that $T' \leq T$ is the number of non-silence frames. The expression $ph(t) \notin \mathrm{SIL}$ indicates removal of silence frames and s is the starting value for the inner summation which is either equal to 0 or 1.

4.7.1.7 Band aperiodicity distortion (BAPD) An aperiodicity in speech waveform is stated as the power ratio of speech signal to the aperiodic element of the signal. BAPD is measured over linearly spaced band aperiodicity coefficients between synthesized voice and ground truth similarly like MCD.

4.7.1.8 Global distortion At each frame, the arithmetic mean of measured distortion is known as global distortion (D) (Loizou 2011) and it is measured as

$$D = \frac{1}{M}\sum_{k=0}^{M-1} d\left(x_k, \bar{x}_k\right) \quad (19)$$

Here, M denotes total frames, $d(x_k, \bar{x}_k)$ is the distance between synthesized and reference sound signals in the k^{th} frame. The distance can be Bark distortion measure (BSD), log-likelihood ratio (LLR) or Itakura-Saito measure.

4.7.2 Subjective testing metrics

4.7.2.1 Mean opinion score (MOS) It is the subjective measure to access the quality of synthesized speech in which listeners are asked to rate the test waveform on the scale of 0 to 5 with 0 means the low unsatisfactory speech quality and 5 indicates the speech quality close to natural human speech (Loizou 2011). This method of MOS is recommended by IEEE subcommittee on subjective methods (Ludovic et al. 2006). The questionnaire is prepared addressing various components related to measure the speech quality (Viswanathan and Viswanathan 2005) such as overall impression (perceived quality of speech to listener), listening effort (effort required to understand the message), pronunciation (perceived anomaly in pronunciation of words), speaking rate (average rate of delivery), articulation (sounds distinguishability) and voice pleasantness (voice characteristics).

4.7.2.2 Preference test In this test, listeners are presented with reference signal and the synthesized signal and they are asked their preference of the signal. Test of significance (p-value test) has been carried out using z-score method, t-test or chi-square method at some threshold value (confidence level). If the obtained p-value is greater than threshold then null hypothesis is accepted otherwise, rejected.

5 More challenging TTS

The naturalness of synthesized speech not only measures the resemblance to human voice but also the way how artificially generated voice express the emotions in the sentence. Therefore, this section demonstrates the challenging TTS to make this paper more competitive and useful for future research.

5.1 Expressive TTS

To improve the naturalness of synthetic speech and enrich the user's experience of using TTS, it is required to integrate expression of emotions into synthesized speech (Schroder 2001). Three aspects of the speech "what, how and who to say" needs to be addressed in TTS. "What to say" is addressed by the front end of TTS by writing sentence, "who to say" is managed by collecting voice data of a person and then training an acoustic model to mimic the speaker's voice. The "how to say" aspect of speech generation is controlled by correct prosody, rate of speech tone and emotion of the articulated speech (Mu et al. 2021). Thus, expressive TTS (E-TTS) addressed this third aspect of speech and are capable to synthesized voice with diversity of speaking styles to express different emotions (happy, sad, affection etc.) and intentions (lodging a complaint, suggesting an advice, making a request etc.) (Wu et al. 2021). Several researches have been carried out to implement E-TTS from more than a decade. One of the approaches employed the prosodic rules for generating synthetic sound to express emotions of joy, sadness, anger, fear, surprise, boredom etc. These rules are pull out by different authors from literature and some by examining their own corpus based on the level and range of f0 during its rises and fall, tempo of voice and its loudness. Table 4 below listed parameter settings for some of the expressions (Schroder 2001). These rules when employed in diaphone synthesis, unit selection based concatenative

Table 4 Parameter setting for different types of emotions for emotion synthesis

Type of emotion	Joy	Sadness	Anger	Fear	Surprise	Boredom
Speaker	German	American English	British English	German	American English	Dutch
F0 Mean	+50%	0, reference line = −1, less final lowering = −5	+ 10 Hz	+150%	0, reference line = −8	65 Hz
F0 range	+50%	−5, steeper accent shape = +6	+9 s.t	+20%	+8, rising slope = +10, accent shape = +5	4 s.t
Tempo	+30%	−10, more fluent pause = −5, hesitation pauses = +10	+30wpm	+30%	+4, less fluent pauses = −5, hesitation pauses = −10	150%

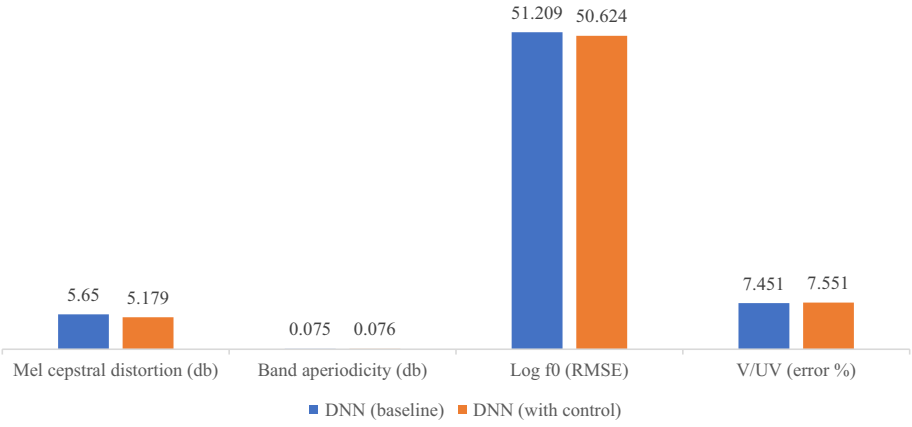


Fig. 18 Objective metrics obtained using and without controlled SPPS

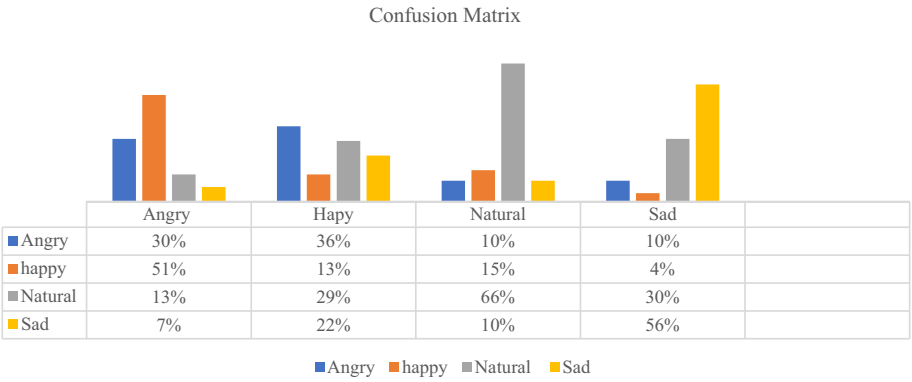


Fig. 19 Confusion matrix for diverse emotions

synthesis needs different models for each emotion. It is possible only when sustainable amount of data is available for each expression of emotion.

To work with small corpus, the adaption of singe model to synthesize emotive speech was first introduced in SPPS by Yamagishi et al. (2004) using HMMs and then more complex form of adaption using Cluster Adaptive Training (CAT) (well suited for DNN acoustic modeling) is demonstrated for implementing E-TTS (Chen et al. 2014a). These approaches (Hodari et al. 2018; Lee et al. 2017) uses the categorical code and dimensional values to synthesize emotive voice. Unlike natural TTS, where input to synthesizer is only the written text, E-TTS requires one additional module that use the control vector of labeled emotions (Lee et al. 2017). Such controllable SPPS systems use the external data to train emotional recognition system which is then used to predict labels for diverse emotions if natural version of sentence is previously existed. For novel sentence, a manual specification of control vector is given (Hodari et al. 2018). This predicted label is given along with written text as input to synthesizer to synthesize emotive speech waveforms. Interactive emotional dyadic motion capture dataset (IEMOCAP) which contains speech of 1.5 h is used. Utterances were marked for categorical and continuous emotions and accuracy of

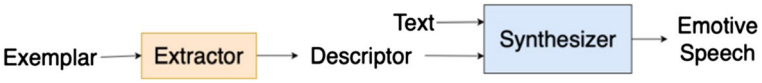


Fig. 20 E-TTS using emotion descriptor obtained using emotion exemplar

emotional classification achieved was 62.9% using neural network of two private layers (20 units) and one shared layer (200 units). Figures 18 and 19 below (Hodari et al. 2018) shows the result obtained for 2 synthesized voices using controlled SPPS and without controlled SPPS and the confusion matrix for the accuracy of predicted labels.

However, E-TTS based on categorical code and dimensional values are not able to express complex emotions and suffers from annotation ambiguity respectively. Therefore, Wu et al. (2021) propose a technique of using emotion exemplar to extract emotional descriptor which are then used with input text to generate the emotive voice at the time of synthesis as described in Fig. 20.

While synthesizing voice using emotion descriptor, which features should be used for emotion exemplar, how they are mapped to emotion descriptor and how to incorporate them to generate emotive speech needs to be addressed. Features are represented by examining the spectrogram of the speech. For example, in neutral sentence of utterance "you" a flat contour is seen in spectrogram and salient pitch rise in the beginning of utterance in emphatic "you". These spectrogram features are mapped using capsule and residual neural network into emotion descriptor. Capsule neural networks are used to acquire spatial data from these spectrogram and residual neural networks obtains the contrasting features in neutral and emotive sentences for the similar written input. These descriptors are then augmented with input text to generate emotive voice using sequence to sequence architecture. Figure 21 below represents the subjective and objective evaluation results of four systems—emotion code vector (EC), emotion probability values (EP), logit-based descriptor (EP) and automatically derived latent emotions (EA) TTS.

In the similar context, Wang et al. (2018b) proposed a global style token (GST) network to generate emotive speech. In this, GST network is learned using speech data having various styles represented using style token. For each style token, a reference signal is used as a guidance process to weight each token. Further, to assign weights to these style token, Kwon et al. (2019a) introduced methods based on controlled weight (CW) by inspecting the emotions' distribution in emotional vector space. For the distribution of each emotion. This is further improved by Um et al. (2020) who introduced embedding vector to calculate

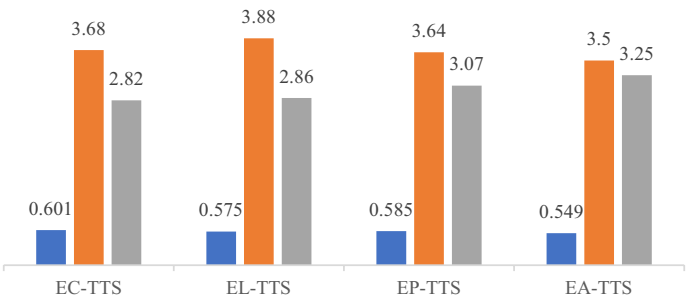


Fig. 21 Comparison of systems based on objective and subjective evaluation

inter-to-intra emotion distance (I2I) for each emotional category (Kwon et al. 2019b) by reducing the distance from target emotion category. They also proposed the method of spread-aware I2I to control the intensity of expressing emotions. Instead of using linear interpolation, spread-aware I2I used the interpolation method that slowly changes the weight value assigned to each emotion style in emotion embedding vectors. With this it is made possible to synthesize emotive voice with different types of emotions of weak anger, gloomy sad or strong happiness. Mellotron (Valle et al. 2020) further included the fundamental frequency (f0) value to this. Zhang et al. (2019a, b) used the variational autoencoder (VAE) to train the network on latent variables representing the style features.

5.2 Multi-speaker, multi-lingual TTS

Single speaker TTS based on deep learning methods are easily extended to support multiple voices using multitask learning by replication of the output layer for individual speaker in training corpus. As shown in Fig. 22, each speaker has its own-trained regression output layer whereas all share the hidden layers of acoustic neural network model (Fan et al. 2015).

This model is well suited when limited training data is available by freezing the hidden layer and only changing the regression layer but it expands linearly with the number of training speakers. Multi-lingual support is easily extended to multi-speaker TTS using IPA training. Liu and Mak (2019) worked on the cross-lingual TTS system using the shared phonemes for the input and speaker, language, stress and tone embedding and trained it on English and Mandarin. The Griffin-Lim (GL) vocoder was initially used to synthesize native voice but the same was unable to synthesize accented sound. Therefore, this work was further improved by same authors in 2020 (Liu and Mak 2020) by introducing the accent control to generated accented speech and native speech control in case produced voice is not native to speaker. For synthesizing more natural speech, authors used the phoneme embedding as an alternative to character embedding. Here also shared phoneme set is used by mapping Mandarin and Cantonese phonemes to ARPABET (ARPAbet, phonetic transcription codes by Advanced Research Project Agency) except the phonemes 'j', 'q' and 'x' as no good mapping is found in ARPABET. WaveNet, an audio synthesizer, with

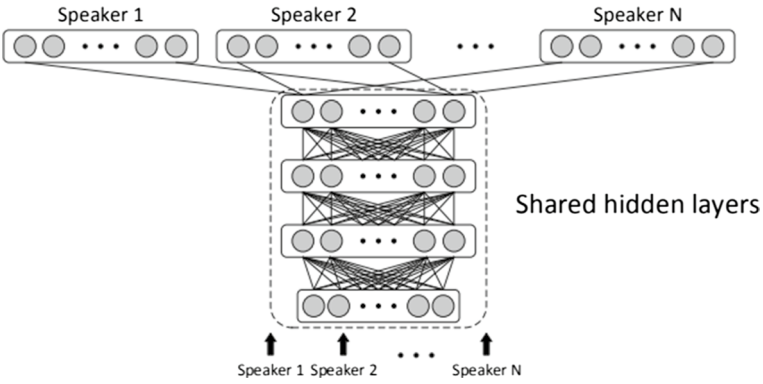


Fig. 22 Multi-speaker DNN architecture (Fan et al. 2015)

Table 5 Naturalness and Intelligibility MOS values with 95% confidence value (Liu and Mak 2020)

Target language (TL)/ source language (SL)	Metric	Cantonese	English	Mandarin
Cantonese	Intelligibility MOS—native)	4.50 ± 0.11	3.28 ± 0.23	3.84 ± 0.20
English	Intelligibility MOS—accented	–	2.05 ± 0.27	1.70 ± 0.15
	Naturalness MOS—native	4.28 ± 0.18	2.92 ± 0.23	3.51 ± 0.21
	Naturalness MOS—accented	–	2.13 ± 0.17	1.54 ± 0.09
	Intelligibility MOS—native)	4.07 ± 0.15	4.54 ± 0.07	3.78 ± 0.14
	Intelligibility MOS—accented	3.09 ± 0.14	–	2.14 ± 0.17
	Naturalness MOS—native	3.79 ± 0.20	4.30 ± 0.13	3.45 ± 0.20
	Naturalness MOS—accented	2.91 ± 0.20	–	2.05 ± 0.17
Mandarin	Intelligibility MOS—native)	4.43 ± 0.09	4.17 ± 0.15	4.42 ± 0.15
	Intelligibility MOS—accented	2.26 ± 0.21	2.57 ± 0.20	–
	Naturalness MOS—native	4.24 ± 0.12	3.70 ± 0.15	4.32 ± 0.18
	Naturalness MOS—accented	2.12 ± 0.35	2.30 ± 0.19	–

Table 6 Synthesis time and MOS comparison of ParaNet and Deep Voice 3

TTS	Synthesis time (ms)	MOS Score
Deep Voice 3 + WaveNet	$181.8 + 5 \times 10^5$	4.09 ± 0.26
ParaNet + WaveNet	$3.9 + 5 \times 10^5$	4.01 ± 0.24
Deep Voice 3 + WaveVAE (variational autoencoder)	$181.8 + 64.9$	3.70 ± 0.29
ParaNet + WaveVAE	$3.9 + 64.9$	3.31 ± 0.32

20 convolution layers is trained using tone and stress embedding to synthesize native and accent-controlled speech. The results are demonstrated in Table 5.

Multi-speaker speech can be easily synthesized by introducing speaker embedding vector (Gatys el al. 2016; Goodfellow et al. 2014) in the encoding unit. Researchers such as Jia et al. (2018), Arik et al. (2018) and Nachmani et al. (2018) presented the speaker encoder in well-known TTS Tacotron 2, Deep Voice 3 and VoiceLoop (Taigman et al. 2017) respectively which encode the information of speaker in the reference speech. This speaker embedding vector can be easily obtained from the speech corpus of target speaker. The speech corpus required to have recording from a different number of speakers (Table 6).

6 State of arts TTS employing deep learning

Recently, Speech synthesizer founded on deep learning use autoregressive model (AR) (Oord et al. 2016) to synthesis high fidelity waveforms. These models generally composed of pipelining in which first mel-spectrogram is generated from written text and then, vocoder is used to artificially generates speech waveform from the generated mel-spectrogram. Such structures only require audio and corresponding transcript as a training data (Wei et al. 2017; Shen et al. 2018a, b). However, in these autoregressive models, synthesis is relatively slow because of the lack of parallelism at training and synthesis time and generated speech is not robust, words are skipped and sometimes repeated. Now, non-autoregressive

(NAR) models are also proposed in literature for parallel waveform generation (Oord et al. 2016; Kumar et al. 2019; Ping et al. 2018; Binkowski et al. 2020) which possesses higher synthesis rate but lower accuracy as compared to their AR counterparts. Knowledge distillation (Oord et al. 2017; Ren et al. 2019), word mapping (Guo et al. 2019), source target alignment constraints (Gu et al. 2017), duration prediction (Ren et al. 2019) are some techniques that are used to lessen the deprivation of accuracy.

6.1 TTS based on autoregressive models

6.1.1 WaveNet

Introduced in 2016 by Google DeepMind, WaveNet (Oord et al. 2016, 2017; Zhao et al. 2018)) facilitated the end-to-end speech synthesis. It is fully autoregressive and probabilistic model. The synthesized sound generated by WaveNet is natural sounding. It has deep convolution neural network (CNN) model which is trained using speech recording. It models the probability distribution of the audio waveforms based on the previously produced speech waveforms. During training phase, WaveNet is trained using huge recording of speech samples to determine the original structures of speech samples. The synthetic speech is then generated using one sample at a time from scratch with 24,000 samples per second. The generated voice is varied by gender, language and accent. It supports multiple languages spoken in the world like Arabic, Bengali, Chinese (Hong Kong), English (Australia), English (India) and many more. Even some languages support multiple voices. Figure 9 shows the Mean Option Score (MOS) rating for different voices generated using WaveNet. MOS rating expresses the opinion of people concerning the quality of speech. The test results shown in Fig. 23 for US English voice produced using WaveNet got average MOS rating of 4.1 on the scale of 1–5 which is 20% improved than the standard voice. This reduces the gap between natural human voice and synthesized voice by 70%.

The new improved version of WaveNet runs on Google Cloud Tensor Processing Unit (TPU) infrastructure, generated speech in less than 50 ms hence producing 1000 times faster than the original model. But still error commenced at front-end effects the synthesis process.

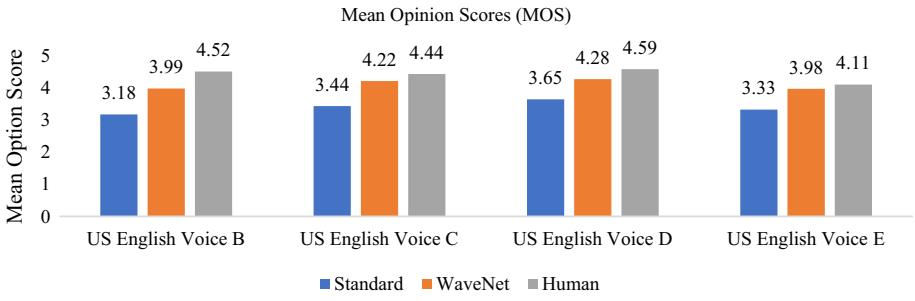


Fig. 23 Wavenet MOS rating for US English Voice

6.1.2 Tacotron

Launched in 2017, Tacotron uses a generative model similar to WaveNet for end-to-end speech synthesis (Wang et al. 2017a, 2017b) Given a < text, audio > pair, Tacotron model used random initialization for training (Chung et al. 2018). It synthesized speech using seq2seq model that map the linguistic text to the mel- spectrogram. Since mel-spectrogram doesn't consist of phase information therefore, missing phase information is estimated using Griffin–Lim algorithm (Grifiin and Lim et al. 1984). Tacotron is applied to almost every language because it used characters as a basic unit to synthesized speech. In addition, it has MOS rating of 3.82 on the scale of 1–5 for US English and synthesized speech at frame level which makes it faster than the WaveNet or other TTS which synthesized speech at sample level. Further, it can be extended easily based on the availability of transcript acoustic data hence it does not involve any phoneme level alignment.

Tacotron 2 Tacotron-2 is enhanced version of Tacotron which is released by Google in 2018. It combines both CNN and RNN for synthesizing high quality natural speech. The model generates the mel-spectogram from the text and then modified WaveNet vocoder is used to generate time-domain speech waveforms from frequency based mel-spectrograms. In comparison to original Tacotron version, Tacotron-2 used simplified building blocks such as LSTM and CNN for encoding and decoding. Figure 24 below shows the MOS rating of various well know TTS approaches with Tactron 2.

Tacotron yielded simpler pipelining by employing seq2seq framework with the extension capability which makes it appropriate for various challenging tasks like voice cloning (Arik et al. 2018), adapting the system to new speaker (Chen et al. 2019).

6.1.3 Deep-voice

Deep voice TTS (Arik et al. 2017a) first presented in 2017 by Baidu Silicon Valley Artificial Intelligence Lab and then improved versions Deep Voice 2 and Deep Voice 3 are released soon in 2017. The Deep Voice TTS is based on the neural networks which comprises of segmentation model, model to change grapheme-to-phoneme, models for

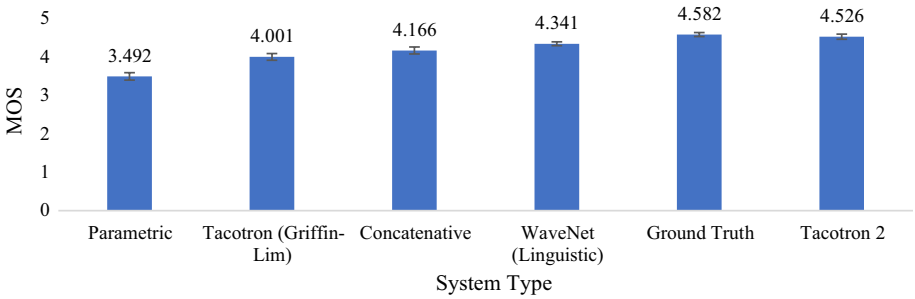


Fig. 24 Comparison of MOS rating of Tacotron 2 with other models

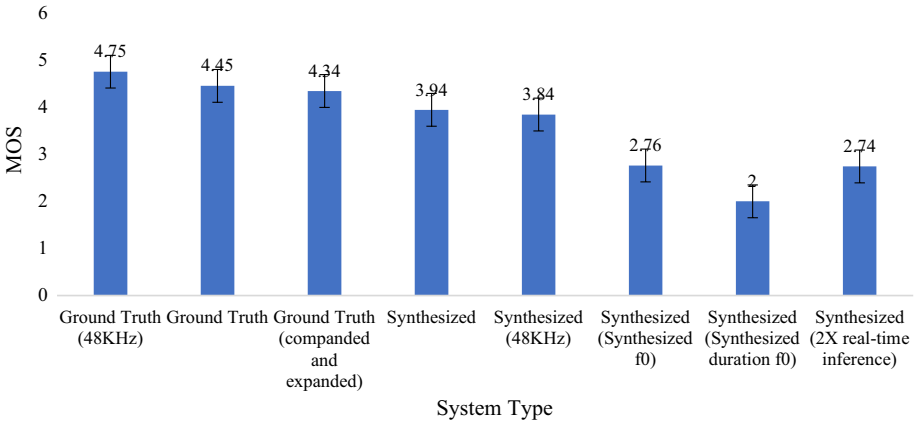


Fig. 25 Comparison of Deep Voice synthesized sound on MOS rating and 95% confidence interval with other models

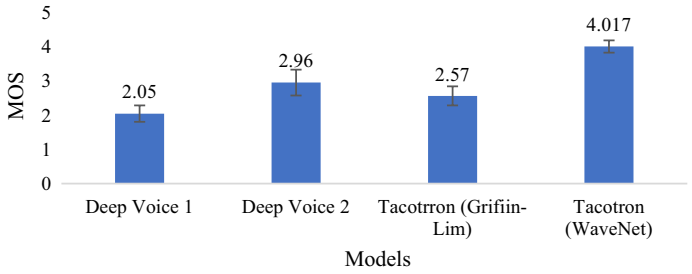


Fig. 26 Comparison of Deep Voice 2 with other well-known models

predicting phoneme durations and its f0 value, variant of a WaveNet model that needs fewer hyperparameters to synthesized audio. The Fig. 25 below shows the comparison of synthesized sound on MOS rating and 95% confidence interval with eight models. The different models vary in size where 'l' is the number of layers, 'r' is the count of residual channels (dimension of the hidden state of every layer), and 's' is the count of skip channels used in each model. Model size for systems 4–7 is l = 40, r = 64, s = 256 and for last model is l = 40, r = 64, s = 256.

Deep Voice 2 (Arik et al. 2017b) is the variant of Deep Voice 1 that is capable of generating several voices using a single model. The architecture of Deep Voice 2 is analogous to Deep Voice 1 with separate modules to predict phoneme duration and f0. In Deep Voice 1, single module jointly determines the phoneme duration and its f0. Further, the segmentation model of Deep voice 2 use convolution RNN with batch normalization. Figure 26 below shows that comparison of Deep Voice 2 synthesized sound with Deep Voice 1 and Tacotron using Griffin-Lim and WaveNet based vocoder.

The third variant Deep voice 3 (Wei et al. 2017) use full convolution neural network that is learned on LibriSpeechASR dataset. The architecture of Deep Voice 3 consists of three main components encoder to convert linguistic features to internal representation, decoder

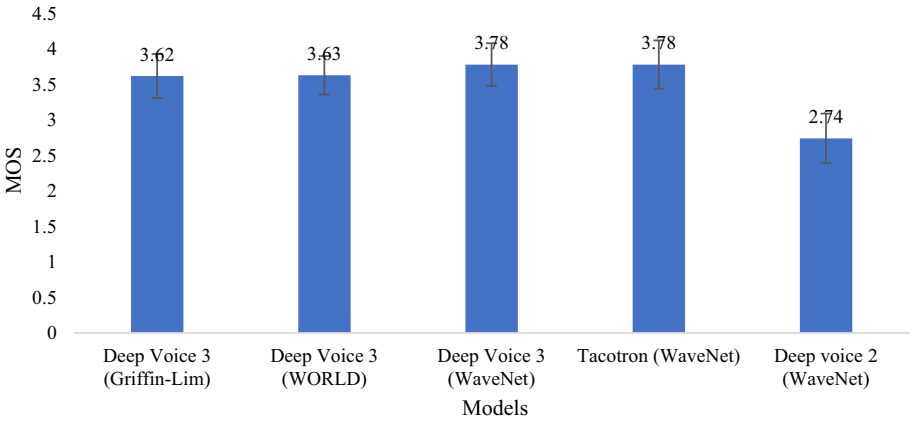


Fig. 27 Comparison of Deep Voice 3 with other state of art models

used to decode learned internal representations and converter to convert learned representations to vocoder parameters. Figure 27 below represents the comparison of Deep Voice 3 with other state of art models on MOS rating.

6.2 TTS based on non-autoregressive model

6.2.1 ParaNet

ParaNet (Peng et al. 2020) is a fully convolutional attention based non autoregressive TTS which generates speech waveform from text in one feed-forward pass. It iterates in layer-by-layer fashion and use attention distillation from autoregressive encoder and repeatedly refine the alignment between text and mel-spectrogram. It has similar encoder as of autoregressive Deep Voice 3 except the parallel decoder which is composed of firstly an attention block for positional encoding followed by fully convolution block and one more attention block. For improving performance, it also predicts log-linear spectrograms in addition with log-mel spectrogram. When compared with autoregressive model (Wei et al. 2017), it is found to be 254 times faster than autoregressive models for real-time synthesis and speed up to 46.7 times as shown inTable 6.

6.2.2 Parallel Tacotron

Parallel Tacotron (Isaac et al. 2021) is a lightweight convolution non-autoregressive TTS model whose training and synthesis is performed parallelly using variational autoencoder

Table 7 Comparison of synthesized sounds using Parallel Tacotron and Parallel Tacotron 2

	Natural speech	Parallel Tacotron	Parallel Tacotron 2
MOS	4.49 ± 0.05	4.42 ± 0.05	4.46 ± 0.05
Preference	Reference	-0.08 ± 0.09	0.01 ± 0.09

which competently captures the local contexts. Its variant Parallel Tacotron 2 (Isaac et al. 2021) is a neural network non-autoregressive end-to-end text-to-speech synthesis model whose duration model is based on the attention distillation and iterative refinement using soft dynamic time wrapping to reduce spectrogram loss. During training, it doesn't rely on either supervised duration signals or teacher forcing of target duration instead it uses fully differentiable duration model to propagate error gradients to the network as a result it synthesizes sound more naturally and fast compared to Parallel Tacotron. Table 7 below shows the comparison of both parallel Tacotrons on MOS rating and preference.

6.2.3 FastSpeech

Proposed by research team at Microsoft and Zhejiang university, FastSpeech TTS (Ren et al. 2019) is able to produce mel-spectrogram 270×and voice by 38×by generating mel-spectrogram parallelly. It predicts the phoneme duration by extracting the attention from teacher model used in autoregressive models using mean square error (MSE) loss and have six feed forward transformation blocks for converting phoneme sequence to mel-spectrogram sequence. Any mismatch of length found in both sequence is bridged using length regulator. However, it doesnot accurately determine duration using teacher model and generated mel-spectrogram also bears information loss as these get simplified using teacher model which reduces the end quality of speech. These limitations are addressed in its variant FastSpeech 2 (Ren et al. 2020). The information loss, here, is prevented by training FastSpeech 2 directly from the ground-truth target instead using simplified data from teacher model. This model determines the values of duration, pitch and energy from target speech and use the continuous wavelet transformation method to translate the pitch contour into pitch spectrogram. It attains the 3×speed up over its original version FastSpeech by using simplified training pipelining and high-quality voice.

6.2.4 Parallel WaveNet

The Parallel WaveNet (Oord et al. 2017) uses parallel feed forward network that use probability Density Distillation method for training. This model generates high-fidelity speech waveforms used online by Google Assistant application. It can synthesize several speaker voices with varying accents. The table gives the comparison of MOS rating on English and Japanese voices using Parametric, concatenative and Distilled WaveNet. The Fig. 28 shows the significantly improved MOS score to be achieved by Distilled WaveNet.

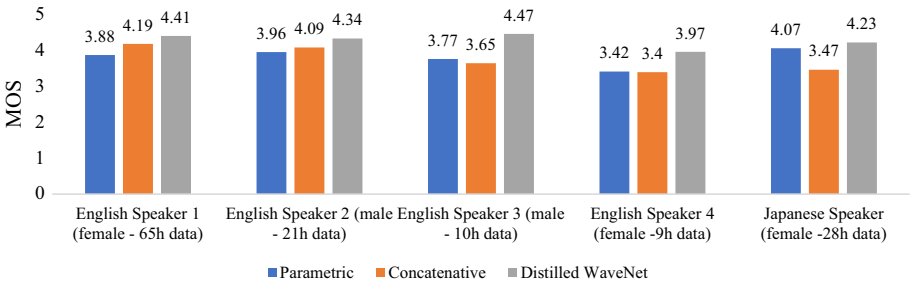


Fig. 28 Comparison of MOS score on English and Japanese with multi-speaker distilled Wavenets

Table 8 Various speech corpus used for training acoustic model

Languages	Speech corpus	Corpus description	Availability
English speech corpus	VCTK	48 h of speech from 109 native speakers with character labeling at sampling rate of 48 kHz	https://datashare.is.ed.ac.uk/handle/10283/2119 (free)
	LibriTTS	585 h of speech form 2,456 speakers with characters labeling and contextual information at sampling rate of 24 kHz	http://www.openslr.org/60/ (free)
	LJ Speech English	24 h of speech from single speaker with character and phoneme labeling at sampling rate of 22.05 kHz	https://keithito.com/LJ-Speech-Dataset/ (free)
	CMU ARCTIC corpus	7 h of speech from 7 speakers with character labeling at sampling rate of 16 kHz	http://www.festvox.org/cmu_arctic/ (free)
	Blizzard Challenge 2011	16.6 h of speech from single speaker with character labeling at sampling rate of 16 kHz	http://www.cstr.ed.ac.uk/projects/blizzard/ (Registered Users)
	Blizzard Challenge 2013	300 6 h of speech from single speaker with character labeling at sampling rate of 44.1 kHz	http://www.cstr.ed.ac.uk/projects/blizzard/ (Registered Users)
Mandarin	CSMSC	12 h of speech from single speaker with Pinyin, rhythm and phoneme boundary at sampling rate of 48 kHz	https://www.data-baker.com/open_source.html (non-commercial use)
	AISHELL-3	85 h of speech from 218 speakers with characters and pinyin labeling at sampling rate of 44.1 kHz	http://www.aishelltech.com/aishell_3 (academic research)
	DiDiSpeech	6000 h of speech from 800 speakers with characters and standardized pinyin labeling at sampling rate of 44.1 kHz	https://outreach.didichuxing.com/research/opendata/ (free)

Table 8 (continued)

Languages	Speech corpus	Corpus description	Availability
Multi-lingual	CSS10	Single speaker speech corpus with character labeling at sampling rate of 22 kHz Supported Languages- Chinese, Dutch, French, Finnish, Japanese, Hungarian, Greek, German, Russian and Spanish	https://github.com/Kyubyong/CSS10.10 (free)
	Common Voice	9,283 h of speech with character encoding; Supports 60 languages	https://commonvoice.mozilla.org/.11 (free)
	Spoken Wikipedia Corpus (SWC-2017)	English—182 h of speech German—395 speakers 249 h of speech from 339 speakers Dutch—79 h of speech from 145 speakers Phoneme level alignment with sampling rate of	https://nats.gitlab.io/swc/ (free)
Danish	NST Danish Speech Synthesis	4108 utterances form single speaker with word level segmentation at sampling rate of 44 kHz	https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-21/ (free)
Swedish	NST Swedish Dictation	4108 utterances form 195 speakers with word level segmentation at sampling rate of 22 kHz	https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-17/ (free)
Norwegian	NST Norwegian Speech Synthesis	5363 utterances form single speaker with word level segmentation at sampling rate of 44 kHz	https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-15/ (free)
Indian Languages	Hindi Raw Speech Corpus	121 h of speech from 488 speakers at sampling rate of 48 kHz	http://data.ldcil.org/hindi-raw-speech-corpus
	Gujarati Raw Speech Corpus	57:17:08 h of speech from 96 female and 108 male speakers	http://data.ldcil.org/gujarati-raw-speech-corpus

7 Speech corpus used for training

For implementing TTS, a speech corpus of high quality with labeled data is needed. Some available speech corpora are summarized below in Table 8

8 Conclusion

This paper gives the review of various approaches used to synthesized voice from the written text. The strategy for selection of papers for the current issue is based on the five metrics such as google scholar h5 index, SCIMago journal ranking, journal quartile (Q1-Q4), number of citations of the article and year of publication. This article encapsulates the contemporary as well as conventional technologies that are used to build acoustic model of TTS in detail. It further classifies the methods based on its characteristics and also provides the advantages of using each stated technology with possible research gaps. Articulatory and formant synthesis rule-based methods used formerly although generates highly intelligible sound but synthesized voice sounds robotic, therefore, gained no much popularity. After that, concatenative approaches such as unit-based and diaphone-based speech synthesis methods, even though generates the natural speech, need larger speech corpus from single speaker. In addition, reading style voice is generated and sound is not continuous at frame boundaries. Compared with these conventional approaches to speech synthesis, the SSPS especially, deep learning-based methods provides more accurate acoustic modeling by synthesizing a voice from its parametric representation. These models are best to model the complex and nonlinear relationships between the input linguistic features and the output speech parameters. Further, the use of bidirectional LSTM in RNN architecture enable to mode the temporal features and DGPs helps to alleviate the over-fitting problems in DNNs to great extent and produced speech by utilizing multi-speaker speech corpus.

The article also reviews a methodology to integrate emotions to the speech synthesizers to generate emotive and expressive voice. Multi-lingual and multi-speakers TTS are discussed along with here. In addition, various objective metrics such as Itakura-Saito measure, RMSE of log f0, Gross Pith Error, V/UV error rate, Correlation Coefficient, Mel Cepstral distortion, Band Aperiodicity Distortion, Global Distortion and subjective metrics such as Mean Opinion Score and Preference test used to access the quality of the synthesized speech are examined.

Further, state-of-art TTS employing the model of autoregressive and non-autoregressive are reviewed in this survey. On the one hand autoregressive TTS are able to regenerate speech of high quality, on the other non-autoregressive TTS are able to synthesis speech at high rate. These well-known TTS like WaveNet, Tacotron, Tacortron 2, Deep Voice 1, Deep Voice 2, Deep Voice 3, parallel WaveNet, parallel Tacotron, FastSpeech have employed the deep learning methods to train their acoustic model and the MOS rating for them to be observed is 4.1, 4.001(Griffin-Lim Vocoder), 4.526, 2.05, 2.96, 3.78, 4.71, 4.46 ± 0.05, 3.84 ± 0.08 respectively.

Acknowledgements This piece of research work in supported by IK Gujral Punjab Technical University, Kapurthala, Punjab, India.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

Achanta S, Gangashetty S (2017) Deep Elman recurrent neural networks for statistical parametric speech synthesis. Speech Commun 93:31–42. https://doi.org/10.21437/Interspeech.2015-266

Alias F, Sevillano X, Socoró JC, Gonzalvo X (2008) Towards high-quality next-generation text-to-speech synthesis: A multidomain approach by automatic domain classification. IEEE Trans Audio Speech Lang Process 16(7):1340–1354. https://doi.org/10.1109/TASL.2008.925145

Arik S, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y, Li X, Miller J, Raiman J, Sengupta S, Shoeybi M (2017a) Deep Voice: real-time neural text-to-speech. https://arxiv.org/abs/1702.07825

Arik S, Diamos G, Gibiansky A, Miller J, Peng K, Ping W, Raiman J, Zhou Y (2017b) Deep Voice 2: multi-speaker neural text-to-speech. https://arxiv.org/abs/1705.08947

Arik SO, Chen J, Peng K, Ping W, Zhou Y (2018). Neural voice cloning with a few samples. In: Neural information processing system. https://doi.org/10.48550/arXiv.1802.06006

Babacan O, Drugman T, D'Alessandro N, Henrich N, Dutoit T (2013) A comparative study of pitch extraction algorithms on a large variety of singing sounds. In: IEEE international conference on acoustics, speech and signal processing, pp 7815–7819. https://doi.org/10.1109/ICASSP.2013.6639185

Biagetti G, Crippa P, Falaschetti L (2018) HMM speech synthesis based on MDCT representation. Int J Speech Technol 21(4):1045–1055. https://doi.org/10.1007/s10772-018-09571-9

Binkowski M, Donahue J, Dieleman S, Clark A, Elsen ´ E, Casagrande N, Cobo LC, Simonyan (2020) High fidelity speech synthesis with adversarial networks. In: International conference on learning representation. https://doi.org/10.48550/arXiv.1909.11646

Botha GR, Barnard E (2012) Factors that affect the accuracy of text-based language identification. Comput Speech Lang 26(5):307–320. https://doi.org/10.1016/j.csl.2012.01.004

Chai L, Du J, Liu QF, Lee CH (2019) Using generalized gaussian distributions to improve regression error modeling for deep learning-based speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 27(12):1919–1931. https://doi.org/10.1109/TASLP.2019.2935803

Chen L, Gales MJF, Braunschweiler N, Akamine M, Knill K (2014a) Integrated expression prediction and speech synthesis from text. IEEE J Select Top Signal Process 8(2):323–335. https://doi.org/10.1109/JSTSP.2013.2294938

Chen LH, Ling ZH, Liu LJ, Dai LR (2014b) Voice conversion using deep neural networks with layer-wise generative training. IEEE/ACM Trans Audio Speech Lang Process 22(12):1859–1872. https://doi.org/10.1109/TASLP.2014.2353991

Chen LH, Raitio T, Valentini BC, Ling ZH, Yamagishi LH (2015) A deep generative architecture for post-filtering in statistical parametric speech synthesis. IEEE/ACM Trans Audio Speech Lang Process (TASLP) 23(11):2003–2014. https://doi.org/10.1109/TASLP.2015.2461448

Chen Y, Assael Y, Shillingford B, Budden D, Reed S, Zen H, Wang Q, Cobo LC, Trask A, Laurie B et al. (2019). Sample efficient adaptive text-to-speech. https://doi.org/10.48550/arXiv.1809.10460

Chung YA, Wang Y, Hsu WN, Zhang Y, Skerry-Ryan RJ (2018) Semi-supervised training for improving data efficiency in end-to-end speech synthesis. https://doi.org/10.48550/arXiv.1808.10128

Coelho LP, Braga D, Dias MS, Mateo CG (2013) On the development of an automatic voice pleasantness classification and intensity estimation system. Comput Speech Lang 27(1):75–88. https://doi.org/10.1016/j.csl.2012.01.006

Cutajar K, Bonilla EV, Michiardi P, Filippone M (2017) Random feature expansions for deep Gaussian processes. In: Proceeding of international conference on machine learning, pp 884–893. https://arxiv.org/abs/1610.04386

Damianou AC, Lawrence N (2013) Deep gaussian processes. In: Proceeding of international conference on artificial intelligence statistic, pp 207–215. https://doi.org/10.48550/arXiv.1211.0358

Dong S, Wang P, Abbas K (2021) A survey on deep learning and its applications. Comput Sci Rev 40:100379. https://doi.org/10.1016/j.cosrev.2021.100379

Drugman, T, Wilfart G, Dutoit T (2009) A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. https://arxiv.org/abs/2001.00842

Fan Y, Qian Y, Soong F, He L (2015) Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4475–4479. https://doi.org/10.1109/ICASSP.2015.7178817

Fukada T, Tokuda K, Kobayashi T, Imai S (1992) An adaptive algorithm for Mel-cepstral analysis of speech. In. Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 137–140. https://doi.org/10.1109/ICASSP.1992.225953

Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2414–2423. https://doi.org/10.1109/CVPR.2016.265

Ghahabi O, Hernando J (2018) Restricted boltzmann machines for vector representation of speech in speaker recognition. Comput Speech Lang 47:16–19. https://doi.org/10.1016/j.csl.2017.06.007
Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. https://doi.org/10.48550/arXiv.1406.2661
Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Boca Raton
Griffin D, Lim J (1984) Signal estimation from modified short-time Fourier transform. IEEE Trans Acoust Speech Signal Process 32(2):236–243. https://doi.org/10.1109/TASSP.1984.1164317
Gu J, Bradbury J, Xiong C, Li VO, Socher R(2017). Non-autoregressive neural machine translation. https://doi.org/10.48550/arXiv.1711.02281
Gujarathi P, Patil SR (2021) Review on unit selection-based concatenation approach in text to speech synthesis system. In: Cybernetics, cognition and machine learning applications, Springer, Singapore, pp 191–202. https://doi.org/10.1007/978-981-33-6691-6_22
Guo J, Tan X, He D, Qin T, Xu L, Liu TY (2019) Non-autoregressive neural machine translation with enhanced decoder input. Proc AAAI Conf Artif Intell 33:3723–3730. https://doi.org/10.48550/arXiv.1812.09664
Hinton G, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527
Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T, Kingsbury B (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. Signal Process Mag 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597
Hodari Z, Watts O, Ronanki S, King S (2018) Learning interpretable control dimensions for speech synthesis by using external data. In: Proceeding of INTERSPEECH 2018, pp 32–36. https://doi.org/10.21437/Interspeech.2018-2075
Hojo N, Ijima Y, Mizuno H (2018) DNN-based speech synthesis using speaker codes. IEICE Trans Inf Syst 101:462–472. https://doi.org/10.1587/transinf.2017EDP7165
Hu YJ, Ling ZH (2016) DBN-based spectral feature representation for statistical parametric speech synthesis. IEEE Signal Process Lett 23(3):321–332. https://doi.org/10.1109/LSP.2016.2516032
Huang WC, Hayashi T, Wu YC, Kameoka H, Toda T (2021) Pretraining techniques for sequence-to-sequence voice conversion. IEEE/ACM Trans Audio Speech Lang Process 29:745–755. https://doi.org/10.1109/TASLP.2021.3049336
Ijima Y, Miyazaki N, Mizuno H, Sakauchi S (2015) Statistical model training technique based on speaker clustering approach for HMM-based speech synthesis. Speech Commun 71:50–61. https://doi.org/10.1016/j.specom.2015.04.003
Isaac E, Heiga Z, Jonathan S, Zhang Y, Jia Y, Ron W, Yonghui W (2020). Parallel Tacotron: non-autoregressive and controllable TTS. In: IEEE international conference on acoustics, speech and signal processing, pp 5709–5713. https://doi.org/10.48550/arXiv.2010.11439
Isaac E, Heiga Z, Jonathan S & Zhang Y, Jia Y & Ryan RJ & Yonghui W (2021) Parallel Tacotron 2: A non-autoregressive neural TTS model with differentiable duration modeling. https://doi.org/10.48550/arXiv.2103.14574
Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, Chen Z, Nguyen P, Pang R, Moreno IL, et al. (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. https://doi.org/10.48550/arXiv.1806.04558
Jinyin C, Ye L, Ming Z (2021) MASS: multi-task anthropomorphic speech synthesis framework. Comput Speech Lang 70:1012–1043. https://doi.org/10.48550/arXiv.2105.04124
Juang BH (1984) On using the Itakura-Saito measures for speech coder performance evaluation. At&t Bell Lab Tech J 63(8):1477–1498. https://doi.org/10.1002/j.1538-7305.1984.tb00047.x
Kang S, Qian X, Meng H (2013) Multi-distribution deep belief network for speech synthesis. In: IEEE international conference on acoustics, speech and signal processing, pp 8012–8016. https://doi.org/10.1109/ICASSP.2013.6639225
Karabetsos S, Tsiakoulis P, Chalamandaris A, Raptis S (2010) One-class classification for spectral join cost calculation in unit selection speech synthesis. Signal Process Lett 17(8):746–749. https://doi.org/10.1109/LSP.2010.2053357
Kawahara H, Katayose H, Cheveigné A, Patterson R (1999) Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In: Proceedings of EUROSPEECH, pp 2781–2784. https://doi.org/10.21437/Eurospeech.1999-613
Kayte SN, Mundada M, Kayte C (2015) A review of unit selection speech synthesis. Int J Adv Res Comput Sci Softw Eng 5(10):475–479

Khorinphan C, Phansamdaeng S, Saiyod S (2014) Thai speech synthesis with emotional tone: Based on formant synthesis for home robot. In: 2014 third ICT international student project conference (ICT-ISPC), IEEE, pp 111–114. https://doi.org/10.1109/ICT-ISPC.2014.6923230

Kolen JF, Kermer SC (2001) Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In: A field guide to dynamical recurrent neural, pp 237–244. https://doi.org/10.1109/9780470544037.ch14

Kominek J, Schultz T, Black AW (2008) Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In: A field guide to dynamical recurrent networks, pp 63–68. https://doi.org/10.1109/9780470544037.ch14

Koriyama T, Kobayashi T (2019) Statistical parametric speech synthesis using deep gaussian processes. IEEE/ACM Trans Audio Speech Lang Process 27(5):948–959. https://doi.org/10.1109/TASLP.2019.2905167

Kumar K, Kumar R, de Boissiere T, Gestin L, Teoh WZ, Sotelo J, de Brébisson A, Bengio Y, Courville AC (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. In: Advances in neural information processing systems, pp 14910–14921. https://doi.org/10.48550/arXiv.1910.06711

Kwon O, Song E, Kim JM, Kang HG (2019a) Effective parameter estimation methods for an ExcitNet model in generative text-to-speech systems. https://doi.org/10.48550/arXiv.1905.08486

Kwon O, Jang I, Ahn C, Kang HG (2019b) An effective style token weight control technique for end-to-end emotional speech synthesis. IEEE Signal Process Lett 26(9):1383–1387. https://doi.org/10.1109/LSP.2019.2931673

Latif S, Rana R, Khalifa S, Jurdak R, Qadir J, Schuller BW (2020) Deep representation learning in speech processing: Challenges, recent advances, and future trends. https://doi.org/10.48550/arXiv.2001.00378

Le N, Rathour VS, Yamazaki K, Luu K, Savvides M (2021) Deep reinforcement learning in computer vision: a comprehensive survey. Artif Intell Rev. https://doi.org/10.48550/arXiv.2108.11510

Lee CH, Jung SK, Kang HG (2007) Applying a speaker-dependent speech compression technique to concatenative TTS synthesizers. IEEE Trans Audio Speech Lang Process 15(2):632–640. https://doi.org/10.1109/TASL.2006.876762

Lee Y, Rabiee A, Lee SY (2017) Emotional end-to-end neural speech synthesizer. https://arxiv.org/abs/1711.05447

Li Y, Lee T, Qian Y (2004) Analysis and modeling of F0 contours for Cantonese text-to-speech. ACM Trans Asian Lang Inf Process 3(3):169–180. https://doi.org/10.1145/1037811.1037813

Liang MS, Yang RC, Chiang YC, Lyu DC, Lyu RY (2004) A Taiwanese text-to-speech system with applications to language learning. In: proceedings of the IEEE International Conference on Advanced Learning Technologies, pp, 91–95. https://doi.org/10.1109/ICALT.2004.1357381

Ling ZH, Wu YJ, Wang YPu Ping; Qin, Long; Wang, Ren Hua (2006) USTC System for blizzard challenge 2006: An improved HMM-based speech synthesis method. In: Proceeding of blizzard challenge workshop

Ling ZH, Deng L, Yu D (2013) Modeling spectral envelopes using restricted Boltzmann machines for statistical parametric speech synthesis. IEEE Trans Audio Speech Lang Process 21(10):7825–7829. https://doi.org/10.1109/TASL.2013.2269291

Ling ZH, Kang SY, Zen H, Senior A, Schuster M, Qian XJ, Meng HM, Li D (2015) Deep learning for acoustic modelling in parametric speech generation: a systematic review of existing techniques and future trends. IEEE Signal Process Mag 32(3):35–52. https://doi.org/10.1109/MSP.2014.2359987

Liu ZC, Ling ZH, Dai LR (2018) Statistical parametric speech synthesis using generalized distillation framework. IEEE Signal Process Lett 25(5):695–699. https://doi.org/10.1109/LSP.2018.2819886

Liu Z, Mak B (2019) Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers. https://doi.org/10.48550/arXiv.1911.11601

Liu Z, Mak B (2020) Multi-lingual multi-speaker text-to-speech synthesis for voice cloning with online speaker enrollment. In: proceeding of INTERSPEECH, pp 2932–2936

Loizou PC (2011) Speech quality assessment. Multimedia analysis, processing and communications. Stud Comput Intell 346:623–654

Ludovic M, Berger J, Kastner M (2006) The ITU-T standard for single-ended speech quality assessment. IEEE Trans Audio Speech Lang Process 14(6):1924–1934. https://doi.org/10.1109/TASL.2006.883177

Lukose S, Upadhya S (2017) Text to speech synthesizer-formant synthesis. In: Proceeding of 2017 international conference on nascent technologies in engineering (ICNTE), pp 1–4. https://doi.org/10.1109/ICNTE.2017.7947945

Mametani K, Kato T, Yamamoto S (2019) Investigating context features hidden in end-to-end TTS. In: Proceedings of IEEE the 44th international conference on acoustics, speech and signal processing, Brighton, pp 6920–6924. https://doi.org/10.1109/ICASSP.2019.8683857
Mitsui K, Koriyama T, Saruwatari H (2020) Multi-speaker text-to-speech synthesis using deep Gaussian processes. In: Proceeding of INTERSPEECH 2020. https://doi.org/10.48550/arXiv.2008.02950
Mitsui K, Koriyama T, Saruwatari H (2021) Deep Gaussian process based multi-speaker speech synthesis with latent speaker representation. Speech Commun 132:132–145. https://doi.org/10.1016/j.specom.2021.07.001
Moungsri D, Koriyama T, Kobayashi T (2018) GPR-based Thai speech synthesis using multi-level duration prediction. Speech Commun 99:114–123. https://doi.org/10.1016/j.specom.2018.03.005
Mu Z, Yang X and Dong Y (2021) Review of end-to-end speech synthesis technology based on deep learning. https://doi.org/10.48550/arXiv.2104.09995
Nachmani E, Polyak A, Taigman Y, Wolf L (2018) Fitting new speakers based on a short untranscribed sample. In: International conference on machine learning, PMLR. pp 3683–3691. https://doi.org/10.48550/arXiv.1802.06984
Nakamura K, Hashimoto K, Oura K, Nankaku Y, Tokuda K (2019) Singing voice synthesis based on convolutional neural networks. https://doi.org/10.48550/arXiv.1904.06868
Nakashika T, Yatabe K (2021) Gamma Boltzmann machine for audio modeling. IEEE/ACM Trans Audio Speech Lang Process 29:2591–2605. https://doi.org/10.1109/TASLP.2021.3095656
Nakashika T, Takaki S, Yamagishi J (2019) Complex-valued restricted Boltzmann machine for speaker-dependent speech parameterization from complex spectra. IEEE/ACM Trans Audio Speech Lang Process 27(2):244–254. https://doi.org/10.1109/TASLP.2018.2877465
Narendra NP, Rao KS (2017) Generation of creaky voice for improving the quality of HMM-based speech synthesis. Comput Speech Lang 42:38–58. https://doi.org/10.1016/j.csl.2016.08.002
Nazir O and Malik A (2021) Deep learning end to end speech synthesis: a review. In: international conference on secure cyber computing and communications (ICSCCC), pp 66–71. https://doi.org/10.1109/ICSCCC51823.2021.9478125
Norvig P, Russel JS (2020) Artificial intelligence: a modern approach.
Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016). WaveNet: a generative model for raw audio. https://arxiv.org/abs/1609.03499
Oord A, Li Y, Babuschkin I, Simonyan Karen, Vinyals O, Kavukcuoglu K, Driessche G, Lockhart E, Cobo L, Stimberg F, Casagrande N, Grewe D, Noury S, Dieleman S, Elsen E, Kalchbrenner N, Zen H, Graves A, King H, Hassabis D (2017) Parallel WaveNet: fast high-fidelity speech synthesis. https://doi.org/10.48550/arXiv.1711.10433
Panda S, Nayak A (2015) An efficient model for text-to-speech synthesis in Indian languages. Int J Speech Technol 18(3):305–315. https://doi.org/10.1007/s10772-015-9271-y
Panda SP, Nayak AK, Patnaik S (2015) Text to speech synthesis with an Indian language perspective. Int J Grid Utility Comput 6:170–178. https://doi.org/10.1504/IJGUC.2015.070676
Peng K, Ping W, Song Z, Zhao K (2020) Non-autoregressive neural text-to-speech. In: International conference on machine learning, pp 7586–7598. https://doi.org/10.48550/arXiv.1905.08459
Peter B, Susanne D, Simon S (2019) Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis. INTERSPEECH 2019: 3765–3769. https://doi.org/10.21437/Interspeech.2019-2410
Ping W, Peng K, Chen J (2018). ClariNet: Parallel wave generation in end-to-end text-to-speech. https://doi.org/10.48550/arXiv.1807.07281
Ping W, Peng K, Zhao K, Song Z (2020). WaveFlow: a compact flow-based model for raw audio. In: International conference on machine learning. https://doi.org/10.48550/arXiv.1912.01219
Qian Y, Fan y, Hu W, Soong FK (2014) On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 3829–3833. https://doi.org/10.1109/ICASSP.2014.6854318
Qinsheng D, Jian Z, Lirong W, Lijuan S (2011) Articulatory speech synthesis: a survey. In: Proceeding of 14th IEEE international conference on computational science and engineering, pp 539–542. https://doi.org/10.1109/CSE.2011.95
Rao KS, Narendra P (2019) Source modeling techniques for quality enhancement in statistical parametric speech synthesis. In: Springer briefs in speech technology: studies in speech signal processing, natural language understanding, and machine learning, pp 13–15
Reddy K, Rao S (2017) Robust pitch extraction method for the HMM-based speech synthesis system. IEEE Signal Process Lett 24(8):1133–1137. https://doi.org/10.1109/LSP.2017.2712646

Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, Liu TY (2019) FastSpeech: fast, robust and controllable text to speech. In: Proceeding of advances in neural information processing systems. https://doi.org/10.48550/arXiv.1905.09263

Ren Y, Hu C, Tan X, Qin, T, Zhao S, Zhao Z, Liu T Y (2020). Fastspeech 2: fast and high-quality end-to-end text to speech. https://arxiv.org/abs/2006.04558

Saito Y, Takamichi S, Saruwatari H (2019) Vocoder-free text-to-speech synthesis incorporating generative adversarial networks using low-/multi-frequency STFT amplitude spectra. Comput Speech Lang 58:347–363. https://doi.org/10.1016/j.csl.2019.05.008

Salami R, Laflamme C, Bessette B, Adoul JP (1997) ITU-T G.729 annex a: reduced complexity 8 kbit/s CS-ACELP codec for digital simultaneous voice and data. IEEE Commun Mag 35(9):53–63. https://doi.org/10.1109/35.620526

Sasirekha D, Chandra E (2012) Text to speech: a simple tutorial. Int J Soft Comput Eng (IJSCE) 2(1):275–278

Schmidhuber J (2014) Deep learning in neural networks: an overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schroder M (2001) Emotional speech synthesis: a review. In: 7th European conference on speech communication and technology, pp 561–564. https://doi.org/10.21437/Eurospeech.2001-150

Sharma B, Prasanna SRM (2017) Enhancement of spectral tilt in synthesized speech. IEEE Signal Process Lett 24(4):382–386. https://doi.org/10.1109/LSP.2017.2662805

Sharma P, Abrol V, Sao AK (2018) Reducing footprint of unit selection-based text-to-speech system using compressed sensing and sparse representation. Comput Speech Lang 55:91–208. https://doi.org/10.1016/j.csl.2018.05.003

Shen J, Pang R, Weiss R J, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerry-Ryan R et al. (2018a) Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In: IEEE International conference on acoustics, speech and signal processing. https://doi.org/10.1109/ICASSP.2018a.8461368

Shen J. Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan, RJ, Saurous RA, Agiomvrgiannakis Y, Wu Y (2018b) Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In: Proceeding of IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4779–4783. https://doi.org/10.48550/arXiv.1712.05884

Shinoda K, Watanabe T (2000) MDL-based context-dependent subword modeling for speech recognition. J Acoust Sci Technol 21(2):79–86. https://doi.org/10.1250/ast.21.79

Siddhi D, Verghese JM, Bhavik D (2017) Survey on various methods of text to speech synthesis. Int J Comput Appl 165(6):26–30. https://doi.org/10.5120/ijca2017913891

Sisman B, Yamagishi J, King S, Li H (2021) An overview of voice conversion and its challenges: from statistical modeling to deep learning. IEEE/ACM Trans Audio Speech Lang Process 29:132–157. https://doi.org/10.1109/TASLP.2020.3038524

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the annual conference on neural information processing systems, pp 3104–3112. https://doi.org/10.48550/arXiv.1409.3215

Tabet Y, Boughazi M (2011) Speech synthesis techniques. A survey. In: Proceeding of international workshop on systems, signal processing and their applications, pp 67–70. https://doi.org/10.1109/WOSSPA.2011.5931414

Takamichi S, Toda T, Neubig G, Sakti S, Nakamura S (2014) A Postfilter to modify modulation spectrum in HMM-based speech synthesis. In: Proceeding of international conference of acoustics, speech, and signal processing, pp 290–294. https://doi.org/10.1109/ICASSP.2014.6853604

Taigman Y, Wolf L, Polyak A, Nachmani E (2017) Voiceloop: voice fitting and synthesis via a phonological loop. https://doi.org/10.48550/arXiv.1707.06588

Talkin D (1995) A robust algorithm for pitch tracking (RAPT). Speech coding and synthesis, pp 497–518

Tiomkin S, Malah D, Shechtman S, Kons Z (2011) A hybrid text-to-speech system that combines concatenative and statistical synthesis units. IEEE Trans Audio Speech Lang Process 19(5):1278–1288. https://doi.org/10.1109/TASL.2010.2089679

Tits N, Haddad KE, Dutoit T (2019) Exploring transfer learning for low resource emotional TTS. Intelligent systems and applications. In: proceeding of advances in intelligent systems and computing, pp 53–60. https://doi.org/10.1007/978-3-030-29516-5_5

Toda T (2011) Modeling of speech parameter sequence considering global variance for HMM-based speech synthesis. In: Hidden Markov models, theory and applications, pp 131–150

Toda T, Tokuda K (2007) A Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE transactions on information and systems E90-D (5):816–824. http://dx.doi.org/https://doi.org/10.1093/ietisy/e90-d.5.816

Tokuda K, Kobayashi T, and Imai S (1995) Speech parameter generation from HMM using dynamic features. In: International conference on acoustics, speech, and signal processing, pp 660–663. https://doi.org/10.1109/ICASSP.1995.479684

Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T (2000) Speech parameter generation algorithms for HMM-based speech synthesis. In: Proceeding of international conference of acoustics, speech, and signal processing, pp 1315–1318. https://doi.org/10.1109/ICASSP.2000.861820

Tokuda K, Zen H, Black AW (2002) An HMM-based speech synthesis system applied to English. In: Proceeding of IEEE workshop on speech synthesis, pp 227–230. https://doi.org/10.1109/WSS.2002.1224415

Tokuda K, Nankaku Y, Toda T, Zen H, Yamagishi J, Oura K (2013) Speech synthesis based on hidden Markov models. Proc IEEE 101(5):1234–1252

Um SY, Oh S, Byun K, Jang I, Ahn C, Kang HG (2020) Emotional speech synthesis with rich and granularized control. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7254–7258. https://doi.org/10.1109/ICASSP40776.2020.9053732

Valle R, Li J, Prenger R, Catanzaro B (2020) Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In: International conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6189–6193. https://doi.org/10.48550/arXiv.1910.11997

Viswanathan M, Viswanathan M (2005) Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. Comput Speech Lang 19:55–83. https://doi.org/10.1016/j.csl.2003.12.001

Wang C, Ling Z, Zhang B, Dai L (2008) Multi-layer f0 Modeling for HMM-based speech synthesis. 2008 6th International symposium on Chinese spoken language processing, pp 1–4. https://doi.org/10.1109/CHINSL.2008.ECP.44

Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q (2017a) Tacotron: towards end-to-end speech synthesis. In: Proceeding of INTERSPEECH-2017a, pp 4006–4010. https://doi.org/10.48550/arXiv.1703.10135

Wang Y, Skerry-Ryan, RJ, Stanton D, Wu Y, Weiss R, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R, Saurous R (2017b) Tacotron: A fully end-to-end text-to-speech synthesis model. In: Proceeding of INTERPSEECH-2017b. https://arxiv.org/abs/1703.10135

Wang X, Takaki S, Yamagishi J (2018a) Autoregressive neural f0 model for statistical parametric speech synthesis. IEEE/ACM Trans Audio Speech Lang Process 26(8):1406–1419. https://doi.org/10.1109/TASLP.2018.2828650

Wang Y, Stanton D, Zhang Y, Ryan RS, Battenberg E, Shor J, Xiao Y, Jia Y, Ren F, Saurous RA (2018b) Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis. In: International conference on machine learning. pp 5180–5189

Wang Q, Wang X, Liu W, Chen G (2021) Predicting the Chinese poetry prosodic based on a developed BERT model. In: IEEE 2nd international conference on big data, artificial intelligence and internet of things engineering (ICBAIE), pp 583–586

Wei P, Kainan P, Andrew G, Arik SO, Kannan A, Narang S, Raiman J, Miller J (2017) Deep Voice 3: scaling text-to-speech with convolutional sequence learning. In: International conference on learning representations. https://doi.org/10.48550/arXiv.1710.07654

Wen Z, Li K, Huang Z, Lee CH, Tao J (2018) Improving deep neural network-based speech synthesis through contextual feature parametrization and multi-task learning. J Signal Process Syst 90(7):1025–1037. https://doi.org/10.1007/s11265-017-1293-z

Wu Z, Virtanen T, Kinnunen T, Chng ES, Li H (2013) Exemplar-based unit selection for voice conversion utilizing temporal information. In: Proceeding of INTERSPEECH-2013, pp 3057–3061. https://doi.org/10.21437/Interspeech.2013-667

Wu X, Cao Y, Lu H, Liu S, Kang S, Wu Z, Liu X, Meng H (2021) Exemplar-based emotive speech synthesis. IEEE/ACM Trans Audio, Speech Lang Process 29:874–886. https://doi.org/10.1109/TASLP.2021.3052688

Xie C, Lv J, Li Y, Sang Y (2018) Cross-correlation conditional restricted Boltzmann machines for modeling motion system. Knowl Based Syst 159:259–269. https://doi.org/10.1016/j.knosys.2018.06.026

Yamagishi TM, Kobayashi T (2004) HMM-based expressive speech synthesis - Towards TTS with arbitrary speaking styles and emotions. In: Proceeding of special workshop in Maui (SWIM)

Yamagishi J, Nose T, Zen H, Ling ZH, Toda T, Tokuda K, King S, Renals S (2009) Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Trans Audio Speech Lang Process 17(6):1208–1230

Yang J, Klabjan D (2021) Bayesian active learning for choice models with deep gaussian processes. IEEE Trans Intell Transp Syst 22(2):1080–1092. https://doi.org/10.1109/TITS.2019.2962535

Yin X, Lei M, Qian Y, Soong F, He L, Ling ZH, Dai LR (2015) Modeling f0 trajectories in hierarchically structured deep neural networks. Speech Commun 76:82–92. https://doi.org/10.1016/j.specom.2015.10.007

Yin X, Ling ZH, Hu YJ, Dai LR (2016) Modeling spectral envelopes using deep conditional restricted Boltzmann machines for statistical parametric speech synthesis. In: Proceeding of IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5125–5129. https://doi.org/10.1109/ICASSP.2016.7472654

Yishuang N, Sheng H, Zhiyong W, Chunxiao X, Zhang LJ (2019) A review of deep learning-based speech synthesis. Appl Sci 9(19):40–50. https://doi.org/10.3390/app9194050

Zen H, Sak H (2015) Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis.In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4470–4474

Zen H, Tokuda K, Masuko T, Kobayashi T, Kitamura T (2007) A hidden semi-markov model-based speech synthesis system. IEICE-Trans Inf Syst 90:825–834. https://doi.org/10.1093/ietisy/e90-d.5.825

Zen H, Tokuda K, Black AW (2009) Statistical parametric speech synthesis. Speech Commun 51(11):1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

Zen H, Senior A, Schuster M (2013) Statistical parametric speech synthesis using deep neural networks. In: Proceeding of IEEE international conference on acoustics, speech and signal processing, pp 7962–7966. https://doi.org/10.1109/ICASSP.2013.6639215

Zen H, Dang V, Clark R, Zhang Y, Weiss R, Jia Y, Chen Z, Wu Y (2019) LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In: Proceeding of INTERSPEECH, pp 1526–1530. https://doi.org/10.48550/arXiv.1904.02882

Zhang Y, Weiss RJ, Zen H, Wu Y, Chen Z, Skerry-Ryan RJ, Jia Y, Rosenberg A, Ramabhadra B (2019a) Learning to speak fluently in a foreign language: multilingual speech synthesis and cross-language voice cloning. In: the proceeding of INTERSPEECH 2019a. https://doi.org/10.48550/arXiv.1907.04448

Zhang YJ, Pan S, He L, Ling ZH (2019b) Learning latent representations for style control and transfer in end-to-end speech synthesis. In: international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 6945–6949. https://doi.org/10.1109/ICASSP.2019b.8683623

Zhao Y, Takaki S, Luong HT, Yamagishi J, Saito D, Minematsu N (2018) Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder. IEEE Access 6:60478–60488. https://doi.org/10.1109/ACCESS.2018.2872060

Zhou X, Ling ZH, Dai LR (2021) UnitNet: a sequence-to-sequence acoustic model for concatenative speech synthesis. IEEE/ACM Trans Audio Speech Lang Process 29:2643–2655. https://doi.org/10.1109/TASLP.2021.3093823

Zoughi T, Homayoonpoor M (2018) DBMiP: a pre-training method for information propagation over deep networks. Comput Speech Lang 55:82–100. https://doi.org/10.1016/j.csl.2018.10.001