Digital Medicine

OPEN

A review of the text-to-speech synthesizer for human robot interaction for patients with Alzheimer's disease

Junxiao Yu[1,2], Yihao Yao[2], Rui Feng[2], Tao Liang[2], Wei Wang[2,*], Jianqing Li[2,*]

ABSTRACT

With the rapid growth of eldering process worldwide, the number of people with mild cognitive impairment (MCI) has also been largely increased. To ease the problem that not all the patients get diagnosed and treated properly in time, intelligent robot that additionally equipped with cognitive rehabilitation functions are widely researched and gradually applied to either clinics or families. Speech interaction acts as an indispensable part in human robot interaction (HRI) process, speech quality used during which directly affects the HRI efficiency and users' experience. Studies indicate that high-fidelity speeches that can be clearly and naturally expressed are more likely to be received and understood by MCI patients. Also, using voices that are either familiar or appeared to patients, along with positive expression, largely improves emotional accompany and mental consolation for them, which enhance the cognitive rehabilitation process. The real-time voice synthesizer provides sufficient technical support for the development of cognitive robots, which show significance for families, societies, and clinics. This article reviews the research status of the development of text-to-speech (TTS) synthesizers, including the state-of-arts expressive TTS and voice cloning models. In addition, this paper pays attention to the current challenges and prospects of cognitive rehabilitation robots.

Key words: Mild cognitive impairment, cognitive rehabilitation robot, deep learning, expressive speech synthesizer

INTRODUCTION

Alzheimer's disease (AD) is an irreversible neurodegenerative disorder and expressed as memory loss, cognitive decline, and emotional instability at the early stage. Despite continuous advancements in medical technology, there is still no definitive cure for AD, and the progression of the disease can only be delayed through medication and cognitive rehabilitation training. Cognitive rehabilitation training can help patients maintain cognitive function and slow down the decline process, making it widely used in the rehabilitation treatment of AD patients. However, with the worsening aging population worldwide, an increasing number of AD patients are unable to receive timely diagnosis and treatment. With the rapid development of artificial intelligence and robotics technology, intelligent interactive robots with cognitive rehabilitation or companionship functions have been extensively researched and gradually put into use. The core of interactive cognitive rehabilitation robots is voice interaction, which enables direct and explicit communication of information and emotional expression. The key to voice interaction lies in speech synthesis, and the quality of this technology determines the accuracy of information delivery by the robot, directly impacting the user's interactive experience. Based on the reminiscence therapy,[1] people are more likely to remember things when listening to their familiar sounds. Also, people with AD often get emotional. Using soft and cheerful tones and speaking style help them to calm down.[2] Therefore, sounds come from a cognitive rehabilitation robot is important during human-robot interaction process. For AD patients, clear, accurate, and natural pronunciation allows them to efficiently obtain important information during the interaction process. Using positive, optimistic, and cheerful interactive voices can stimulate the enthusiasm of AD patients for interaction and have a soothing effect on their restless and anxious emotions. Based on these theories, this article introduces and analyzes two key technologies in voice interaction: text-to-speech synthesis technology and emotion-based speech synthesis technology.

SPEECH SYNTHESIZER

Speech synthesis technology is an indispensable and critical component of human-machine interaction systems. High-

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[1]The Affiliated Changzhou Second People's Hospital of Nanjing Medical University, Changzhou Second People's Hospital, Changzhou Medical Center, Nanjing Medical University, Changzhou, Jiangsu Province, China, [2]Jiangsu Province Engineering Research Center of Smart Wearable and Rehabilitation Devices, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, Jiangsu Province, China

***Address for correspondence:** Professor Wei Wang, Jiangsu Province Engineering Research Center of Smart Wearable and Rehabilitation Devices, School of Biomedical Engineering and Informatics, Nanjing Medical University, 140 Hanzhong Rd, Gulou, Nanjing, Jiangsu 210029, China. E-mail: bmeww@njmu.edu.cn; https://orcid.org/0000-0002-9702-6316

Professor Jianqing Li, Jiangsu Province Engineering Research Center of Smart Wearable and Rehabilitation Devices, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, Jiangsu Province, China. E-mail: jqli@nju.edu.cn; https://orcid.org/0000-0002-3524-8933

Digit Med 2023;9:e00011

Received: 10 July 2023; **Accepted:** 23 October 2023

https://doi.org/10.1097/DM-2023-00011

quality speech synthesis enables users to convey and receive information in a more intuitive and efficient manner. With the rapid development and advancement of computer science and artificial intelligence, speech synthesis technology has received extensive attention and research. Text-to-speech (TTS), also known as text-to-voice synthesis, is a key aspect of speech synthesis that aims to convert textual content into sound signals that are perceptible to the human auditory system. Speech synthesis technology is applied in various scenarios, such as assisting patients in understanding training content and providing guidance and question-and-answer interactions during cognitive rehabilitation training. TTS models can offer several specific benefits in Alzheimer's care and fit within the broader healthcare context by enhancing communication, cognitive stimulation, and overall quality of life for individuals with AD, such us improving communication, acting as a reminder and assistance for medication schedules, helping AD patients reduce social isolation and offering emotion accompany.

Text to speech synthesis

Speech synthesis is the most challenging aspect to achieve in human-machine interaction. Generating high-quality and high-fidelity speech signals requires the use of complex and precise models, and model training requires extensive support from speech data. Looking back at the history of speech synthesis development, as early as 1997, Möbius *et al.*[3] proposed a speech synthesis model based on a pipeline structure. As shown in Figure 1, this model consists of multiple modules, each trained separately, and collaborates to complete the speech synthesis task. Based on the traditional pipeline design of speech synthesis, models such as Hidden Markov Models [4] (HMMs) and Gaussian Mixture Models have been successively proposed, which are statistical parametric speech synthesis (SPSS) techniques.[5] These models predict and generate speech based on spectrum analysis. However, the traditional pipeline-based speech models suffer from the limitations of high complexity, low synthesis efficiency, and poor speech synthesis quality due to the individual training of each module and the accumulation of errors in each module.

With the rise of the deep learning wave, end-to-end speech synthesis based on Deep Neural Networks (DNNs) has gradually become a focal point of research for scholars. Although research on speech synthesis technology has a history of several decades, DNN-based speech synthesis technology has produced numerous high-quality research outcomes and classic model architectures in just the past decade. DNN-based speech synthesis models can more effectively capture the hidden features of speech and the correlation between text and speech, resulting in more natural and fluent speech.

WaveNet[6] is an autoregressive speech synthesis model based on PixelCNN,[7] proposed by the DeepMind in 2016. It uses Dilated Causal Convolutions (DCC) to capture long-term temporal dependencies in sound and predicts speech data based on the original signal. Although WaveNet has performed well in many text-to-speech applications, its synthesis speed is slow and not suitable for real-time speech synthesis scenarios. Therefore, building upon WaveNet, Mehri *et al.*[8] introduced the SampleRNN model, which is a highly close-to-end speech generation model that uses hierarchical recurrent neural network (RNN) to generate audio sample by sample in the bottom recurrent layers. SampleRNN is capable of capturing potential variations across extremely long sequence spans and performs well in handling dependencies between sequences. Following SampleRNN, Sotelo *et al.*[9] proposed the Char2Wave model, which consists of a reader and a SampleRNN-based vocoder. It can generate high-fidelity speech without requiring manual alignment adjustments. However, strictly speaking, Char2Wave cannot be considered a truly end-to-end model because the modules still need separate training during text front-end processing.

End-to-End TTS

Tacotron is a deep learning-based speech synthesis model proposed by Wang *et al.*[10] from the Google team in 2017. As the first end-to-end model, Tacotron consists of a text encoder, an attention-based decoder, and a vocoder, shown in Figure 2. Tacotron learns the mapping between text and speech using content-based attention mechanism, predicts mel spectrograms with the encoder, and converts them into speech signals using a vocoder based on the Griffin-Lim algorithm. Although the Tacotron model achieved breakthrough results in speech synthesis, it has a complex model structure, high computational complexity, long training time, low efficiency, and requires a large amount of training data to ensure the generalization performance of the model. The pure DNN approach ignores the

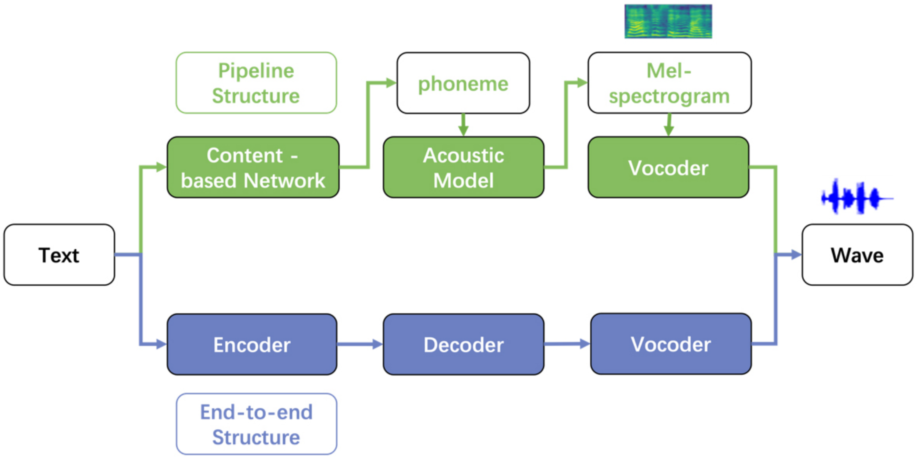

Figure 1. Flow diagram of pipeline and end-to-end TTS models. TTS, text-to-speech.

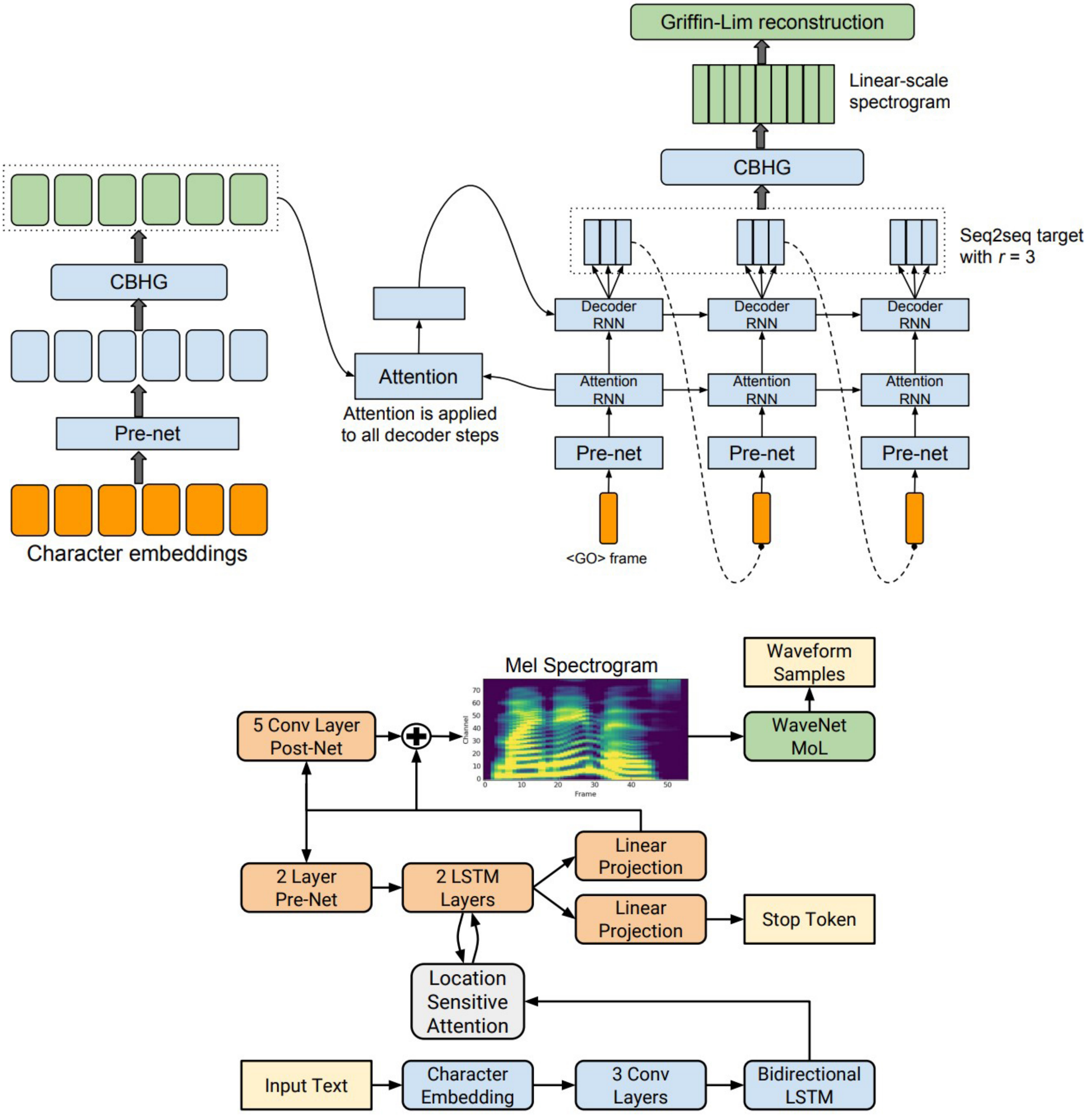

Figure 2. Schematic diagram of Tacotron (left) and Tacotron2 (right) model.

continuity of speech, and the Tacotron model based on DNN samples each frame of the speech signal independently during the training process, resulting in a decrease in the coherence of synthesized speech.

Recurrent Neural Networks (RNNs) compensate for the inability of DNN models to accurately align text and speech by capturing and propagating hidden states over time. Therefore, Shen *et al.*[11] optimized Tacotron and proposed Tacotron2. Tacotron2 simplifies the complex CBHG and High-Way modules in Tacotron1 by combining Convolutional Neural Networks (CNNs) with Long-Short Term Memory (LSTM) recurrent layers, greatly reducing the model parameters and speeding up model training and inference. At the same time, Tacotron2 uses the Local Sensitive Attention (LSA) mechanism to enhance the model's ability to capture the correlation between contexts, improving the robustness of the attention mechanism. The optimized model has shown significant improvements in both training efficiency and speech synthesis quality. Tacotron2 is considered one of the most classic models in text-to-speech (TTS) and has been continuously improved and optimized, giving rise to many excellent model algorithms.

Non-RNN TTS

Although Tacotron2 has achieved excellent performance in speech synthesis, the use of a large number of RNNs in the model results in slow training and inference speed. Therefore, Ping *et al*.[12,25,26] proposed the DeepVoice speech synthesizer, which replaces RNN with a network structure called Residual Gated Convolution (RGC) that effectively captures the correlation between text contexts. DeepVoice uses non-causal and causal CNNs as the encoder and decoder of the model, significantly accelerating the training and inference speed. Meanwhile, Tachibana *et al*.[27] built upon Tacotron and introduced the DCTTS speech synthesizer based on the Text2Mel[28] and SSRN[29] model architectures, where CNNs replace RNNs. Inspired by the Transformer[30] Li *et al*.[13] proposed the non-RNN speech synthesis model Transformer-TTS in 2019. It utilizes multiple sets of multi-head self-attention mechanisms instead of encoders and decoders, ensuring high-quality synthesized speech while greatly improving the speed during training and inference. Zhang *et al*.[31] proposed a non-RNN TTS model especially works for mandarin text to speech synthesis. Although the above models have improved computational efficiency through parallelization, they still require autoregressive generation of acoustic features frame by frame during the inference process, which significantly impacts the speed of speech generation. In addition, the drawback of autoregressive frame-by-frame speech generation is that when the model incorrectly predicts a Mel frame spectrum at a certain moment, this error frame will be propagated and continued as a reference frame for the next moment, leading to further erroneous inference.

Non-Autoregression TTS

To address this issue, models such as FastSpeech,[14] SpeedySpeech,[16] ParaNet,[32] and FastPitch[19] have introduced a Teacher Network, replacing the traditional autoregressive alignment mechanism with knowledge distillation. By parallelizing the experimental models and guided by the autoregressive Teacher Network, the networks learn the attention alignment mechanism correctly, ensuring the quality of the synthesized speech. FastSpeech[14] consists of a feed-forward Transformer neural network that can generate frame-level acoustic features guided by a length regulator. To further reduce model parameters and improve speech synthesis speed, models like DeviceTTS,[22] SpeedySpeech,[7] and Parallel Tacotron[17] use simpler Deep Feed-Forward Sequence Memory Networks (DFSMN), residual distillation CNNs, and Light Weight Convolution (LConv) modules instead of the more complex Transformer based on FastSpeech. These models are trained using knowledge distillation. To simplify the training process even further, Kim *et al*.[23] proposed Glow-TTS based on normalized flow, replacing the Transformer with a decoder module based on Glow, enabling parallel processing of Mel spectrograms. Similarly, Miao *et al*.[24] introduced Flow-TTS, which further improved the model's generation speed based on Glow-TTS optimization. In the same year, Donahue *et al*.[18] proposed EATS, a speech synthesis model based on Generative Adversarial Networks (GANs). The speech generation network based on the cyclic GAN trains the generator and discriminator alternately, enhancing the naturalness and fluency of the generated speech. Table 1 provides an overview of the recent developments in TTS models.

Expressive speech synthesizer

Speech synthesis models have been extensively researched and widely applied. However, speech synthesized by simple text-to-speech models still lacks expressive qualities such as intonation variation, rhythm, and naturalness, resulting in a gap compared to human speech. Therefore, emotional speech synthesis has

Table 1

Typical traditional TTS models

Generation type	Acoustic model	Team	Main network	Alignment	Characteristic
Autoregression	WaveNet	Oord *et al*,[6] 2016	CNN	DCC	Very slow during training and inference
	Sample RNN	Mehri *et al*,[8] 2016	RNN	Stratified Sampling	
	Tacotron	Wang *et al*,[10] 2017	CBHG, DNN	Content-based Attention	The first end-to-end TTS
	Char2Wav	Sotelo *et al*,[9] 2017	RNN	GMM Attention	Very slow during training and inference
	Deep Voice3	Ping *et al*,[12] 2018	RGC, CNN	Dot-Attention	Tradition TTS, very steady
	Tacotron2	Shen *et al*,[11] 2018	LSTM, DNN	Local Sensitive Attention	Tradition TTS, very steady
	Transformer-TTS	Li *et al*,[13] 2019	Transformer	Multi-Head Attention	Drop RNN and fasten the speed
Non Autoregression	Fast Speech	Ren *et al*,[14] 2019	Transformer	Multi-Head Attention	Alignment errors may occur when dealing with long sentences
	Fast Speech2	Ren *et al*,[15] 2020	Transformer	MFA	270 times faster than Transformer
	Speedy Speech	Vainer *et al*,[16] 2020	CNN	Knowledge Dilution	The model is steadier and the alignment is more accurate
	Parallel Tacotron	Elias *et al*,[17] 2020	LConv	HMM Alignment	
	EATS	Donahue *et al*,[18] 2020	CNN, GAN	Self-attention	
	Fast Pitch	Lancucki *et al*,[19] 2020	Transformer	Knowledge Dilution	
	Diffusion TTS	Jeong *et al*,[20] 2021	DDPM	Transformer only for encoder	Stable, faster, and less parameters
	VITS	Kim *et al*,[21] 2021	VAE	Monotonic Alignment	More natural sound
Autoregression& Non Autoregression	DeviceTTS	Huang *et al*,[22] 2020	DFSMN, RNN	-	The model is steadier and the alignment is more accurate
Normalization Flow	Glow-TTS	Kim *et al*,[23] 2020	Transformer, Glow	Knowledge Dilution, MAS	Speech duration prediction
	Flow-TTS	Miao *et al*,[24] 2020	Glow	Multi-head Attention	Shorten the speed of the speech duration prediction

TTS, text-to-speech; CNN, convolutional neural networks; RNN, recurrent neural networks; DCC, dilated causal convolutions; CBHG, a building block used in the Tacotron; RGC, residual gated convolution; LSTM, long-short term memory; MFA, Multi-factor authentication; HMM, hidden markov models; GAN, generative adversarial networks; DDPM, Denoising diffusion probabilistic model; VAE, Variational autoencoder; DFSMN, deep feed-forward sequence memory networks; MAS, Monotonic alignment search.

gradually become a focus of research.

Expressive speech synthesis aims to modify voice characteristics such as timbre, pitch, and rhythm to express different emotions in synthesized speech, making it more natural and closer to human speech. The simplest approach to achieve emotional speech expression is by adding a reference encoder to the basic TTS model to assist in controlling the synthesis of speech with different emotional expressions. This can be done through style transfer methods, using pre-trained reference encoders to directly manipulate various speech style parameters. The second approach is to directly use reference audio information as input to the reference encoder and use the style embedding vectors encoded by the reference encoder to guide the model in synthesizing speech that matches the reference audio. Popular reference encoders include speaker encoders, style encoders, and emotion encoders.

This article introduces various speech synthesis models for expressing different emotions, including voice cloning models, speech emotion synthesis, and emotion synthesis.

Expressive speech synthesis models

Although deep learning-based TTS models have achieved remarkable results in speech synthesis, human speech production is a seemingly simple yet complex process. The same content can be expressed differently through variations in intonation, rhythm, and timbre. However, conventional text-based speech synthesis models can only generate speech with a single tone. As a result, researchers have turned their attention to emotion-based speech synthesis models. Among them, Wang *et al.*[33] proposed the Global Style Token (GST) in 2018, which sparked a wave of interest in emotional speech synthesis. GST is a repository of different style embedding vectors for different speakers. By extracting and classifying speech features related to speaking style, it generates style embedding vectors specific to the current speaker's style. These vectors are then used in conjunction with the Tacotron speech synthesis module to train the model and generate emotionally expressive speech matching the desired speaking style. GST has shown excellent performance in speech emotion expression and has been widely used in different TTS models, such as TPCW-GST[34] and Mellotron.[35] Mellotron can not only add additional expression to the synthesized speech by transferring targeting speakers' personalized speaking style, but also able to simulate songs with different voice print. Tacotron2 is a highly versatile model with synthesized speech that exhibits high generalization and fidelity. Therefore, Tacotron2 has become the preferred choice as the framework for emotion speech expression models. Liu *et al.*[36] proposed Tacotron-PL in 2021, which is based on Tacotron2. By regulating frame-level reconstruction loss and discourse-level style reconstruction loss, collectively referred to as perceptual loss (PL), the relationship between speech style and text becomes more closely linked, enhancing the expressive capabilities of synthesized speech. Moon *et al.*[37] introduced MIST-Tacotron in 2022, which leverages image transfer learning to enhance speech style by transferring the mel-spectrogram of reference speech style. Hortal *et al.*[38] combined GAN with Tacotron2, creating the Gantron model for emotional speech expression. Changing from auto-regression to non-auto regression largely accelerate the inference speech, making the real-time speech generation possible. The optimized models achieve outstanding performance in terms of the expressive text to speech synthesis.

Voice cloning models

Highly generalized speech cloning models can synthesize speech with the vocal characteristics of a target speaker. The simplest speech cloning model involves adding a speaker encoder to a TTS model to capture and extract the voice features of the target speaker. Arik *et al.*[39] built a neural network-based speech cloning model based on DeepVoice3 in 2018. The model utilizes speaker adaptation and speaker encoding methods to achieve speech cloning. In the same year, Nachmani *et al.*[40] extended the VoiceLoop[41] framework by incorporating speaker encoding, enabling the cloning of speech with high similarity to a reference speech based on a short segment of the target speaker's speech. Skerry-Ryan *et al.*[42] proposed speech cloning based on prosody transfer, using embedded vectors based on speaker prosody to control the synthesized speech in Tacotron. With the introduction of the highly stable and high-quality speech synthesis model Tacotron2, Ye *et al.*[43] built the SV2TTS model based on Skerry-Ryan's work, which allows for multi-speaker voice cloning using the Tacotron2 framework. The model includes a speaker encoder that extracts the voice features of the reference speaker, generating speaker embedding vectors with the speaker's vocal characteristics. Through transfer learning, the TTS model is guided to perform voice cloning. This approach, which involves adding additional speaker encoders for multi-speaker voice cloning, requires a large dataset of diverse speakers. However, the dataset requirements are not stringent, and the presence of background noise or slight distortions does not significantly affect the quality of the synthesized speech. During the inference process, only a short audio segment of the speaker needs to be input to the SV2TTS model to achieve satisfactory voice cloning results. However, this zero-shot-based speech cloning approach, while producing highly similar voices, still exhibits a certain gap in terms of naturalness compared to human speech. When the model needs to generate speech for speakers not present in the training set, the results are not ideal.

To address this issue, Cooper *et al.*[44] proposed the Learnable Dictionary Encoding (LDE) mechanism based on SV2TTS. This mechanism improves the effectiveness of speech cloning by incorporating additional speaker voice feature vectors into the decoder PreNet and attention layers of the Tacotron2 model. Cai *et al.*[45] and Shi *et al.*[46] introduced additional loss values and self-feedback constraint mechanisms to control the difference between speaker embedding vectors and synthesized speech, thereby enhancing the robustness and similarity of the synthesized speech. However, autoregressive-based speech cloning models struggle to achieve efficient speech cloning in terms of training and inference speed. Therefore, in 2022, Xue *et al.*[47] proposed the ECAPA-TDNN clone model based on FastSpeech2. It consists of a speaker encoder based on ECAPA-TDNN, FastSpeech2-based TTS, and a HiFi-GAN[48] vocoder, enabling fast, efficient, and high-fidelity speech cloning. Moreover, the model significantly improves the naturalness and similarity of synthesized speech when performing voice cloning for speakers not encountered during the training process.

Emotional expression models in speech

With the emergence of emotion speech synthesis models, people have also started exploring the synthesis of speech with different emotional expressions. Kim *et al.*[52] proposed the ST-TTS model, which combines natural language techniques to generate emotion style labels based on text content, enabling vivid emotional expression during the text-to-speech process. This style labeling approach is more intuitive and efficient compared to style indices or reference speech-based methods. Additionally, the ST-TTS model utilizes efficient autoregressive training methods, greatly improving the synthesis efficiency of the model. Li *et al.*[51] introduced a multi-adaptation emotional speech synthesis model based on Tacotron2. The model employs a multi-scale reference encoder to extract global style and local features from both text and speech, which are then used as additional information for phoneme sequence expansion in Tacotron2 to achieve speech synthesis with different emotional expressions. The Lei team[53,54] has also conducted extensive research and made significant contributions to emotional speech synthesis. In particular, their work in 2022, called MsEmoTTS,[50] is a multi-scale emotion speech synthesis model that includes a Global Emotion Module (GM), Utterance-level Emotion Module (UM), and Local Emotion Module (LM). These modules are responsible for global emotion classification, utterance-level emotional variation, and phoneme-level emotional intensity control, respectively. MsEmoTTS not only performs well in style transfer based on text and reference speech but also allows for synthesizing emotional speech with different emotions through free control. Table 2 intuitively summarizes the development of emotion speech synthesis models in recent years.

COGNITIVE REHABILITATION ROBOTS

Cognitive rehabilitation robots are rehabilitation assistive devices that combine robot technology with cognitive science theories. They aim to help patients with cognitive impairments recover their cognitive functions or delay the deterioration of cognitive function. With the rapid development of artificial intelligence technology, intelligent cognitive rehabilitation robots are considered to be effective means to alleviate the lack of medical resources, shortage of labor, and high workload of caregivers. Research has shown that compared to virtual agents, the use of physical intelligent robots during therapy is more persuasive, can better motivate users, and achieve better user experience.[55] In addition, robots enhance the effectiveness of cognitive rehabilitation training by increasing patient's social interaction, interaction, and positive emotions.[56]

As early as 2015, Sung *et al.*[57] developed a cognitive rehabilitation robot called PARO, which could communicate and interact with the elderly to improve their mental health and cognitive functions. With the rapid development of artificial intelligence, cognitive rehabilitation robots based on different functions have been designed. In 2018, Paletta *et al.*[58] developed a cognitive rehabilitation robot called Amigo for memory training. In a 2020 experiment, Pino *et al.*[59] used a humanoid intelligent robot NAO to assist patients with moderate cognitive impairment in memory training and achieved excellent results. Patients were able to concentrate better during the training process, exhibited a more positive and optimistic attitude, and showed better performance in memory tasks. Researchers at Cornell University in the United States developed a robot called "KASPAR"[60,61] that can help children with autism spectrum disorders with social training through voice interaction while restoring their cognitive rehabilitation abilities.

SUMMARY

This section introduces the development of speech synthesis technology and speech interaction robots, as well as the current research status in China and abroad. Through literature review, the historical development of text-to-speech, emotional speech expression, voice cloning, and emotion speech synthesis technologies are highlighted. The current mainstream speech synthesis model structures and their synthesis effects are summarized. Furthermore, the existing challenges and future prospects are analyzed.

Declaration

Acknowledgement

None.

Table 2

Expressive TTS models

Attention	Expressive model	Team	Main network	Characteristic
Speech Style Enhancement	GST	Wang *et al*,[33] 2018	Reference Encoder, Tacotron	GST
	TPCW-GST	Stanton *et al*,[34] 2018	GST, Tacotron	Content-based GST
	Mellotron	Valle *et al*,[35] 2019	GST, Tacotron2	Rhythm Modification
	Gantron	Hortal *et al*,[38] 2021	GAN, Tacotron	GAN
	Tacotron-PL	Liu *et al*,[36] 2021	Tacotron, Tacotron2	Perception Loss
	MIST-Tacotron	Moon *et al*,[37] 2022	Tacotron2	Transfer Learning
Multi-Speaker Voice Cloning	Neural Voice Cloning	Arik *et al*,[25] 2018	Speaker encoder, DeepVoice3	Voice Cloning
	Voice Cloning Model	Nachmani *et al*,[40] 2018	Speaker encoder VoiceLoop	Voice Cloning
	Expressive Tacotron	Skerry-Ryan *et al*,[42] 2018	Speaker encoder, Tacotron	Voice Cloning
	SV2TTS	Jia *et al*,[43] 2021	Speaker encoder, Tacotron2	Voice Cloning
	ECAPA-TDNN	Xue *et al*,[49] 2021	FastSpeech2, HiFi-Gan	Voice Cloning
Emotional TTS	MsEmoTTS	Lei *et al*,[50] 2022	BERT, GST, TTS	GM, UM, LM
	Multi-Scale ExpressiveTTS	Li *et al*,[51] 2021	Global/Local Reference Encoder Tacotron2	Multi-Scale
	ST-TTS	Kim *et al*,[52] 2021	Glow-TTS	GST

TTS, text-to-speech; GST, global style token; GAN, generative adversarial networks; GM, global emotion module; BERT, bidirectional encoder representations from transformers; UM, utterance-level emotion module; LM, local emotion module.

Author Contributions

Yu J: Conceptualization, Writing—Original draft preparation; Yao Y: Writing—Reviewing and Editing; Feng R: Paper searching and original draft reviewing; Liang T: Paper searching and original draft reviewing; Wang W: Conceptualization, Supervision; Li J: Supervision, Project administration.

Ethics approval and consent to participate

Not applicable.

Source of Funding

This study is supported by the National Key Research and Development Program of China (No. 2022YFC2405600), NSFC (No. 62071241, No. 62075098, No. 62001240), the Leading-edge Technology and Basic Research Program of Jiangsu (No. BK20192004D), the Key Research and Development Program of Jiangsu (No. BE2022160), and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (No. KYCX23_1982).

Conflict of Interest

Jianqing Li is an Editorial Board Member of the journal. The article was subject to the journal's standard procedures, with peer review handled independently of this editor and his research groups.

Data Availability Statement

No additional data.

REFERENCES

1. Woods B, O'Philbin L, Farrell EM, Spector AE, Orrell M. Reminiscence therapy for dementia. *Cochrane Database Syst Rev*. 2018;3(3):CD001120.
2. Swan K, Hopper M, Wenke R, Jackson C, Till T, Conway E. Speech-Language Pathologist Interventions for Communication in Moderate-Severe Dementia: A Systematic Review. *Am J Speech Lang Pathol*. 2018;27(2):836–852.
3. Möbius B, Sproat R, Santen JPV, *et al*. The Bell Labs German text-to-speech system: an overview. Processing Fifth European Conference on Speech Communication and Techology, Greece, 1997.
4. Eddy SR. Hidden Markov models. *Curr Opin Struct Biol*. 1996;6(3):361–365.
5. Zen H, Tokuda K, Black A. Statistical parametric speech synthesis. *Speech Commun*, 2009;51(11):1039–1064.
6. Oord AVD, Dieleman S, Zen H, *et al*. Wavenet: A generative model for raw audio. 2016. https://doi.org/10.48550/arXiv.1609.03499.
7. Van Den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. The 33rd International Conference on Machine Learning (ICML), New York, 2016.
8. Mehri S, Kumar K, Gulrajani I, *et al*. SampleRNN: An unconditional end-to-end neural audio generation model. 2016. https://doi.org/10.48550/arXiv.1612.07837.
9. Sotelo J, Mehri S, Kumar K, *et al*. Char2wav: End-to-end speech synthesis. the 5th International Conference on Learning Representations (ICLR), Toulon, 2017.
10. Wang Y, Skerry-Ryan R J, Stanton D, *et al*. Tacotron: Towards end-to-end speech synthesis, 2017.
11. Shen J, Pang R, Weiss R J, *et al*. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. The 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, 2018.
12. Ping W, Peng K, Gibiansky A, *et al*. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. 2017. https://doi.org/10.48550/arXiv.1710.07654.
13. Li N, Liu S, Liu Y, *et al*. Neural speech synthesis with transformer network. The 33th AAAI Conference on Artificial Intelligence (AAAI), Hawaiian, 2019.
14. Ren Y, Ruan Y, Tan X, *et al*. Fastspeech: Fast, robust and controllable text to speech. The 33rd International Conference on Neural Information Processing Systems (NIPS). Vancouver, 2019.
15. Ren Y, Hu C, Tan X, *et al*. Fastspeech 2: Fast and high-quality end-to-end text to speech. 2020. https://doi.org/10.48550/arXiv.2006.04558.
16. Vainer J, Dušek O. Speedyspeech: Efficient neural speech synthesis. 2020. https://doi.org/10.48550/arXiv.2008.03802.
17. Elias I, Zen H, Shen J, *et al*. Parallel tacotron: Non-autoregressive and controllable tts. The International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, 2021.
18. Donahue J, Dieleman S, Bińkowski M, *et al*. End-to-end adversarial text-to-speech[J]. arXiv preprint arXiv:2006.03575, 2020.
19. Łańcucki A. Fastpitch: Parallel text-to-speech with pitch prediction. The 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, 2021.
20. Zhang C, Zhang C, Zheng S, *et al*. A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative A. 2023:2303–13336.
21. Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech: International Conference on Machine Learning. 2021.
22. Huang Z, Li H, Lei M. Devicetts: A small-footprint, fast, stable network for on-device text-to-speech. 2020. https://doi.org/10.48550/arXiv.2010.15311.
23. Kim J, Kim S, Kong J, *et al*. Glow-tts: A generative flow for text-to-speech via monotonic alignment search[J]. *Adv Neural Inf Process Syst*. 2020;33:8067–8077.
24. Miao C, Liang S, Chen M, *et al*. Flow-tts: A non-autoregressive network for text to speech based on flow. The 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barchelona, 2020.
25. Arık S Ö, Chrzanowski M, Coates A, *et al*. Deep voice: Real-time neural text-to-speech.T he 34th International Conference on Machine Learning (ICML). Sydney, 2017.
26. Gibiansky A, Arik S, Diamos G, *et al*. Deep voice 2: Multi-speaker neural text-to-speech. The 31st International Conference on Neural Information Processing Systems (NIPS). California, 2017.
27. Tachibana H, Uenoyama K, Aihara S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. The 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)[C], Calgary, 2018.
28. Yuan M, Duan Z. Spoofing Speaker Verification Systems with Deep Multi-speaker Text-to-speech Synthesis. 2019. https://doi.org/10.48550/arXiv.1910.13054.
29. Sheng L, Huang D, Pavlovskiy EN. High-quality speech synthesis using super-resolution mel-spectrogram. 2019. https://doi.org/10.48550/arXiv.1912.01167.
30. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. The 31st International Conference on Neural Information Processing Systems (NIPS). California, 2017.
31. Zhang Y, Deng L, Wang Y. Unified mandarin tts front-end based on distilled bert model. 2020. https://doi.org/10.48550/arXiv.2012.15404.
32. Peng K, Ping W, Song Z, *et al*. Parallel neural text-to-speech: the

International Conference on Learning Representations (ICLR), Addis Ababa, 2019.2019.

33. Wang Y, Stanton D, Zhang Y, *et al*. Style tokens. Unsupervised style modeling, control and transfer in end-to-end speech synthesis: the 35th International Conference on Machine Learning (ICML). Stockholmsmässan, 2018.
34. Stanton D, Wang Y, Skerry-Ryan R J. Predicting expressive speaking style from text in end-to-end speech synthesis. The 2018 IEEE Spoken Language Technology Workshop (SLT). Athens, 2018.
35. Valle R, Li J, Prenger R, *et al*. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, 2020.
36. Liu R, Sisman B, Gao G, *et al*. Expressive tts training with frame and style reconstruction loss. *IEEE/ACM Trans Audio Speech Lang Process*. 2021,29:1806–1818.
37. Moon S, Kim S, Choi Y. MIST-tacotron: end-to-end emotional speech synthesis using mel-spectrogram image style transfer. *IEEE Access*. 2022;10:25455–25463.
38. Hortal E, Alarcia R B. GANtron: Emotional Speech Synthesis with Generative Adversarial Networks. 2021. https://doi.org/10.48550/arXiv.2110.03390.
39. Arik S, Chen J, Peng K, *et al*. Neural voice cloning with a few samples. 2018. https://doi.org/10.48550/arXiv.1802.06006.
40. Nachmani E, Polyak A, Taigman Y, *et al*. Fitting new speakers based on a short untranscribed sample. The 35th International Conference on Machine Learning (ICML)[C], Sweden, 2018.
41. Taigman Y, Wolf L, Polyak A, *et al*. Voiceloop: Voice fitting and synthesis via a phonological loop. 2017. https://doi.org/10.48550/arXiv.1707.06588.
42. Skerry-Ryan R J, Battenberg E, Xiao Y, *et al*. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. The 35th International Conference on Machine Learning. Stockholmsmässan, 2018.
43. Jia Y, Zhang Y, Weiss R, *et al*. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. 2018. https://doi.org/10.48550/arXiv.1806.04558.
44. Cooper E, Lai C, Yasuda Y, *et al*. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. The 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, 2020.
45. Cai Z, Zhang C, Li M. From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint. 2020. https://doi.org/10.48550/arXiv.2005.04587.
46. Shi Y, Bu H, Xu X, *et al*. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. 2020. https://doi.org/10.48550/arXiv.2010.11567.
47. Desplanques B, Thienpondt J, Demuynck K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. 2020. https://doi.org/10.21437/Interspeech.2020–2650.
48. Kong J, Kim J, Bae J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. *Adv Neural Inf Process Syst*. 2020;33:17022–17033.
49. Xue J, Deng Y, Han Y, *et al*. ECAPA-TDNN for Multi-speaker Text-to-speech Synthesis. The 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). Singapore, 2022.
50. Lei Y, Yang S, Wang X, *et al*. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Trans Audio Speech Lang Process*. 2022,30:853–864.
51. Li X, Song C, Li J, *et al*. Towards multi-scale style control for expressive speech synthesis. 2021. https://doi.org/10.48550/arXiv.2104.03521
52. Kim M, Cheon S J, Choi B J, *et al*. Expressive text-to-speech using style tag. 2021. https://doi.org/10.21437/Interspeech.2021–465.
53. Lei Y, Yang S, Xie L. Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis. The 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, 2021.
54. Li T, Wang X, Xie Q, *et al*. Controllable crossspeaker emotion transfer for end-to-end speech synthesis. 2021. https://doi.org/10.48550/arXiv.2011.08679.
55. Li J. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *Int J Hum Comput Stud*. 2015,77:23–37.
56. Siciliano B, Khatib O, Kröger T. Springer handbook of robotics. Springer, 2008.
57. Sung HC, Chang SM, Chin MY, Lee WL. Robot-assisted therapy for improving social interactions and activity participation among institutionalized older adults: a pilot study. *Asia Pac Psychiatry*. 2015;7(1):1–6.
58. Paletta L, Fellner M, Schüssler S, *et al*. AMIGO: Towards social robot based motivation for playful multimodal intervention in dementia: the 11th PErvasive Technologies Related to Assistive Environments Conference. Greece, 2018.
59. Pino O, Palestra G, Trevino R, *et al*. The humanoid robot NAO as trainer in a memory program for elderly people with mild cognitive impairment. *Int J Soc Robot*. 2020;12:21–33.
60. Dautenhahn K, Nehaniv C L, Walters M L, *et al*. KASPAR–a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*. 2009;6(3–4):369–397.
61. Wood LJ, Zaraki A, Robins B, Dautenhahn K. Developing Kaspar: A Humanoid Robot for Children with Autism. *Int J Soc Robot*. 2021;13(3):491–508.

How to cite this article: Yu J, Yao Y, Feng R, *et al*. A Review of the text-to-speech synthesizer for human robot interaction for patients with Alzheimer's disease. *Digit Med*. 2023;9:e00011. doi: 10.1097/DM-2023-00011