A Review of Neural Emotional Speech Synthesis

Kevin Liu

1. Introduction

Text-to-speech (TTS) strives to generate authentic emotional speech based on textual input [19]. With the rapid advancement of machine learning [46] and digital signal processing [12], it becomes easier to imbue synthesized speech with natural emotive qualities. This literature review aims to explore the evolving landscape of emotional TTS shed light on the current state-of-the-art and highlight avenues for future research in the realm of emotional speech synthesis through an examination of key advancements in the field.

In recent years, advancements in deep learning [10, 26] and neural networks have led to significant improvements in the quality and naturalness of synthesized speech [3]. After an overview of neural TTS systems, current emotional solutions will be discussed as well as avenues for future research.

2. A History of Neural TTS

End-to-end TTS is a one-pass approach to speech generation where the entire process of converting text into speech waveform is handled by a single model, without the need for intermediate stages or components. End-to-end models already exist [25, 14], and even achieve human-like performance [40, 41] but the math goes over my head so this literature review will not discuss them in detail. Instead, it focuses on neural TTS systems that employ three main components: a text encoder, a feature extractor, and a vocoder [3].

Each component is responsible for a separate part of the process. The text encoder converts plain text input into hidden-state linguistic or acoustic features [19]. These features are then passed to the acoustic model which generates acoustic features that are then fed to the vocoder [33] which generates the final output waveform.

End-to-end synthesis is a lot more difficult because plaintext and waveforms are such different modalities. The progression to fully end-to-end synthesis models could be illustrated with a few key advancements. First, researchers combined the text analysis and acoustic feature extractor to bypass having to explicitly learn alignment [24]. Then, following the success of Google's PixelCNN [8] models at predicting thousands of pixels rapidly at inference time, researchers at DeepMind released WaveNet [73] which autoregressively modeled output from linguistic features directly. Essentially, the acoustic model and vocoder had been combined into one unit. Finally, Google's Tacotron [60] cleverly uses the Griffin-Lim [52] algorithm to produce waveforms on a reduced/dilated set of linguistic features.

2.1 Encoding Text

The text encoder is the first step of the process of generating speech. One of the earliest encoders was concatenative synthesis where prerecorded speech segments are stored in a database and concatenated to form complete utterances. [36]. Many linguistic optimizations like part-of-speech tagging (whether HMM-based [47] or deep learning approaches [2]) and prosodic analysis are employed along with a final dynamic programming optimization to ensure optimal splitting. These segments can be phonemes, syllables, words, or phrases. Each word, phrase, or phoneme in the input text must directly correspond to a specific, pre-recorded audio segment in the database. If the exact segment is available, it is used; if not, the system likely cannot synthesize the desired speech. TTS using this encoder was inflexible due to the egregious amount of recorded data to produce diverse speech variations.

Following concatenative synthesis, several other forms of textual encoding emerge. Parametric encoding systems like HMM-based [61] synthesis utilize hidden Markov models to statistically represent the sequence of speech sounds and their corresponding acoustic features. Text is converted into a sequence of context-dependent labels. These labels provide detailed contextual information for each phoneme, including phonetic, linguistic, and prosodic contexts. This step is crucial because it allows the HMM synthesis system to model how phonemes should be pronounced in different contexts.

More recently, the foundation of modern neural TTS models can be traced back to the transformer architecture introduced by Vaswani et al. in their seminal paper "Attention is All You Need" [5]. This architecture revolutionized natural language processing tasks by introducing the attention mechanism, which allows the model to focus on different parts of the input sequence when generating output. The attention mechanism enables models to capture longer-range dependencies in the input and map them to corresponding prosodic features in the audio output. In modern TTS systems, variants all using some sort of attention-based encoder architecture are commonplace. Tacotron, for instance, utilizes an encoder-decoder framework with attention to converting text to mel-spectrograms, which are then transformed into waveforms using a vocoder. Many extensions of transformer architecture exist; for instance, TransformerTTS [42] builds upon the work in TacoTron by introducing multi-head self-attention [45] to reduce inference times.

Indeed, self-attention-based encoders like BERT after a bit of fine-tuning perform the best in encoding text to its spoken form [63]. While prosody information is usually handled later down in the pipeline, some studies have shown that such information, duration, pitch [21], and pauses (um and uh) [4] can be predicted directly from phonemes. Another interesting advancement is using a diffusion-based probabilistic model to predict prosody to decrease generalization error [13].

2.2 Input Normalization

There exist a variety of standards that enhance the expressiveness and naturalness of synthesized speech. This can be achieved through various techniques and standards, such as SSML (Speech Synthesis Markup Language) [56] and other methods outlined by the W3C (World Wide Web Consortium), among other strategies. The Speech Synthesis Markup Language (SSML) provides a standardized markup language for describing the prosody and structure of synthesized speech. It includes features for controlling aspects such as pitch, rate, volume, and emphasis, which can be manipulated to convey different emotional states. For example, the <emphasis> indicates emphasis on specific words or phrases and the <break> element inserts pauses of varying lengths. Some TTS systems even incorporate specialized input frameworks specifically designed to encode emotion [6, 56]. More advanced models can utilize sentiment analysis to classify untagged input text into specific emotions (e.g., anger, neutrality, sadness) and then adjust their output to reflect those emotions accurately [17, 21]. In 2022, Microsoft Research released a dataset with “prompt engineered tts” with separate styles and prompt encoders to allow for editable speech synthesis [48].

2.3 Vocoders

The primary function of a vocoder, short for "voice encoder," is to convert acoustic features—such as mel spectrograms or pitch—into a speech waveform that can be played back as audible speech [3]. An acoustic model generates these features from the text, and the vocoder transforms them into the final audio signal, producing intelligible and natural-sounding speech.

While no longer used in modern-day neural network synthesis systems, published in 1997, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [54] was a classic vocoder that offered high-quality speech synthesis by separating the spectral envelope, pitch, and aperiodicity [50]. The spectral envelope captures the timbre of the voice, the fundamental frequency represents the pitch, and the aperiodicity component deals with the noise aspects of the speech signal, such as breathiness and fricatives. It reconstructs the speech waveform by combining these elements and allows for large changes in the pitch and speaking rate of produced waveforms without suffering large losses in speech quality[54]. Another more traditional vocoder was WORLD or “Waveform OverLap-add method based on the Residual Excitation and Linear prediction” which was designed for low latency inference [74].

The first Hidden Markov Model (HMM) based vocoder was in 1999, with Yoshimura et al [53]. However, it was not good because the model’s output acoustic features were oversmoothed among other things [61]. Thus, to better capture relationships that span a longer context window and fix over smoothing, deep neural nets (DNN) [59] and LSTM networks were used in 2014 [64]. In 2016, WaveNet [73], developed by Google DeepMind, was a significant advancement in vocoder technology. Unlike traditional vocoders that operate on handcrafted acoustic features, WaveNet generates raw audio waveforms directly through autoregressive modeling, a process where the model predicts each sample of the waveform sequentially based on previous outputs. This approach allows WaveNet to capture intricate details of speech, resulting in highly natural and realistic output. The architecture of WaveNet consists of deep convolutional neural networks (CNNs) with a dilated causal structure. The term "dilated [38] refers to the arrangement of layers in the network, where the receptive field of each CNN layer increases exponentially. Although it produces high-quality speech, generating high-quality speech with WaveNet requires significant computational resources and time as each time step a forward pass over millions of parameters is being performed [3].

Following WaveNet, WaveGlow [71], by NVIDIA, combines Glow-based generative models [29] with WaveNet-like structures. Glow [30] is a type of normalizing flow model, that learns to map data from a simple distribution to a more complex one, typically a complex distribution like natural images or audio waveforms. Combining normalizing flows with WaveNet-like structures, the goal is to leverage the strengths of both approaches. Normalizing flows provide a flexible and efficient way to model complex distributions, while WaveNet-like structures capture the sequential nature of the data. Glow-TTS [29] was perhaps the most successful flow based model which addressed many alignment issues with autoregressive models like TacoTron. By combining in parallel the Transformer-TTS encoder and the Glow decoder architecture, outputs are generated in parallel which achieves incredibly fast inference.

GANs, or Generative Adversarial Networks, are a class of machine learning models that consist of two paired neural networks, a generator, and a discriminator, trained in opposition to each other to produce realistic data. HiFi-GAN (High-Fidelity Generative Adversarial Network) [31] is a GAN-based vocoder that achieves high-quality speech synthesis with significantly lower computational requirements [68] than WaveNet by employing a generator-discriminator setup. NVIDIA’s BigVGan [7], the result of throwing massive computing resources into a GAN-based vocoder trained on the LibriTTS achieves state-of-the-art results on many zero-shot scenarios.

A final class of neural vocoders that has gained popularity is diffusion models. A diffusion model learns to generate data by reversing a gradual, noisy injection process. It starts with data corrupted by noise and progressively refines it, learning to recover the original data distribution through a series of denoising steps. This approach allows the model to produce high-quality synthetic data by effectively modeling the underlying data distribution and has proven to be incredibly effective in practice [11, 72, 43, 51]. A more thorough comparison of recent neural vocoders was conducted by Govalkar et. al. in 2019 [1].

3. Emotion

Emotion significantly influences how we speak by altering prosodic features [15]. Different emotions may also affect volume, speech rate, voice quality, and articulation. For example, happiness often leads to a higher pitch and varied intonation, while sadness results in slower, monotone speech.

3.1 Emotional Datasets

Emotional voice datasets play a crucial role in advancing research in speech processing, sentiment analysis, and emotional intelligence. These datasets typically consist of audio recordings labeled with various emotional states, such as happiness, sadness, anger, and neutrality, allowing researchers to train and evaluate models for emotion recognition and synthesis tasks. Numerous emotional voice datasets have been compiled and utilized in research for a variety of tasks including speech emotion recognition, multimodal emotion recognition, text-to-speech, and more. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [62] contains acted speech and song recordings by professional actors portraying various emotional states. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [32] dataset consists of spontaneous dyadic interactions, capturing a wide range of emotions in naturalistic settings complete with video for multimodal settings.

3.3 Emotional Synthesis

The traditional approach to emotional speech synthesis typically involves categorizing speech into broad emotional labels such as "happy," "sad," "angry," etc. The first problem that arises with this approach is the mapping issue. Since a single text can be spoken in multiple ways depending on the desired emotional expression, the system ends up predicting an average mel-spectrogram of all possible emotional renditions if the emotion is not specifically encoded as an input. This averaging results in synthesized speech that lacks distinct emotional characteristics and sounds unnatural [3].

There have been a few ways around this. Perhaps the most intuitive but inefficient method involves fine-tuning separate models for each emotion. During the generation phase, the system selects the appropriate model based on the desired emotional output, ensuring that each emotion is represented accurately and naturally [21]. Another approach uses autoencoders to encode speech data into a shared latent embedding space, capturing different emotional expressions of the same sentences by the same speaker. This method is followed by a Text-to-Mel (Text2Mel) and then an SSRN (Super-resolution Network) layer that transforms the mel-spectrogram into a high-resolution spectrogram for the final output [67].

Experiments have been conducted to use deep learning approaches to allow for better inter-emotional transfer during model training with Inuoe et. al [36] combining separate speaker and emotion embeddings to generate output while Liu et. al experiment using an automatic speech recognition ASR model to guide model training [34]. Um et. al propose a novel embedding loss that pushes similar emotions closer together and farther ones apart which shows significantly better results compared to other state-of-the-art emotional TTS systems [18].

However, a finite set of coarse emotion labels leads to somewhat limited and averaged expressions of emotions, lacking the subtle nuances that characterize natural human speech [39, 9]. Additionally, this approach relies heavily on manual labeling, which is cumbersome and often insufficiently detailed to capture emotional subtleties.

To address these limitations, recent work has introduced more fine-grained representations of emotion. One approach operates at the phoneme level, using learned hidden-state representations of emotion strength to describe local emotional details, meaning each phoneme in the speech sequence is associated with a specific emotional intensity [23]. Another innovative method is EmoQ-TTS [16], which synthesizes expressive emotional speech by conditioning phoneme-wise emotion information with fine-grained emotion intensity rendered through distance-based intensity quantization, eliminating the need for human labeling. EmoQ-TTS also allows for manual control of emotional expression by adjusting intensity labels.

4.1 Evaluation

Objective metrics such as log spectral distance (LSD) [20] evaluate the spectral similarity between the generated and actual speech. Voice unvoiced error rate assesses the accuracy of voicing decisions or naturalness [69]. The root mean square error (RMSE) of the fundamental frequency provides a measure of the precision with which pitch is reproduced, influencing the perceived pitch accuracy of the synthesized speech.

Complementing these objective measures are subjective evaluation methods, which capture human perceptions and preferences regarding synthesized speech quality. A/B testing, where listeners compare two audio samples and choose the better one is the most practical benchmark.

4.2 Challenges And Future Directions

Data scarcity is a crucial challenge because emotional TTS systems require large amounts of labeled data representing various emotions, which are hard to obtain due to the limited availability of high-quality datasets with diverse emotional annotations and the complexity of labeling emotions accurately. Cross-linguistic emotional synthesis is particularly challenging because different languages have unique syntactic and prosodic features that influence emotional expression. Additionally, to reiterate the first point, many languages lack extensive emotional speech datasets if at all. The work done in the ZeroResource Speech Challenge is a good starting point for data-scarce scenarios where leveraging a combination of automatic speech recognition, speech emotion recognition, and data augmentation techniques could lead to success [MSLAM, TTS by TTS].

Additionally, enhancing TTS systems to understand and utilize context, like the topic of conversation or the speaker's intent, can significantly improve the naturalness and appropriateness of the synthesized speech. Deploying more energy-efficient models with quicker inference times will make the integration of text-to-speech more seamless in smartphones and other connected devices.

Finally, current TTS systems are increasingly natural at producing well-spoken, formal reading speech. This is a small subset of all human speech and it is critical to develop systems that model more informal, spontaneous speech with pauses, hiccups, and variations in speaking rate.

Future research directions include multimodal emotion recognition and transfer learning specific to emotional TTS. Integrating multimodal data, such as text, audio, facial expressions, and body language, can improve emotion recognition and synthesis in TTS systems by providing richer context and enhancing generalization across different contexts. Transfer learning techniques, such as utilizing pre-trained models on large, general-purpose speech datasets and fine-tuning them on smaller emotional datasets, can mitigate data scarcity and improve model performance.

Optimizing TTS with multispeaker to account for inter-speaker variations such as accents and emotional tones. The system must efficiently process and switch between speakers in real-time, requiring highly optimized model architectures or efficient storage.

4.3 Conclusion

This review summarizes a few remarkable advancements in text-to-speech (TTS) technology, particularly the transition from conventional to neural vocoders and the incorporation of emotional expressiveness into synthesized speech. Neural vocoders like WaveNet, Tacotron, and HiFi-GAN have reshaped speech synthesis by offering high-quality and natural-sounding output. Complemented by sophisticated embedding techniques and deep learning techniques, TTS systems now possess the capability to encode emotions, resulting in more human-like interactions. The future of emotional TTS holds great promise, complemented by emerging technologies and ongoing research efforts to broaden the horizons of state-of-the-art TTS systems.

References

[1] P. Govalkar, J. Fischer, F. Zalkow, and C. Dittmar, “A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction,” presented at the Proc. SSW 2019, 2019, pp. 7–12. doi: 10.21437/SSW.2019-2.

[2] S. Yadav, G. Ketepalli, and P. Ragam, “A Deep Learning Approach to Network Intrusion Detection Using Deep Autoencoder,” *Revue d’Intelligence Artificielle*, vol. 34, pp. 457–463, Sep. 2020, doi: 10.18280/ria.340410.

[3] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A Survey on Neural Speech Synthesis.” arXiv, Jul. 23, 2021. Accessed: May 23, 2024. [Online]. Available: http://arxiv.org/abs/2106.15561

[4] Y. Yan *et al.*, “AdaSpeech 3: Adaptive Text to Speech for Spontaneous Style,” arXiv.org. Accessed: May 24, 2024. [Online]. Available: https://arxiv.org/abs/2107.02530v1

[5] A. Vaswani *et al.*, “Attention Is All You Need.” arXiv, Aug. 01, 2023. doi: 10.48550/arXiv.1706.03762.

[6] J. Betker, “Better speech synthesis through scaling.” arXiv, May 23, 2023. Accessed: May 23, 2024. [Online]. Available: http://arxiv.org/abs/2305.07243

[7] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “BigVGAN: A Universal Neural Vocoder with Large-Scale Training.” arXiv, Feb. 16, 2023. doi: 10.48550/arXiv.2206.04658.

[8] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, “Conditional Image Generation with PixelCNN Decoders.” arXiv, Jun. 18, 2016. doi: 10.48550/arXiv.1606.05328.

[9] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, “CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech,” presented at the Proc. Interspeech 2020, 2020, pp. 4387–4391. doi: 10.21437/Interspeech.2020-1251.

[10] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[11] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “DiffWave: A Versatile Diffusion Model for Audio

Synthesis." arXiv, Mar. 30, 2021. doi: 10.48550/arXiv.2009.09761.

[12]
W. D. Stanley, G. R. Dougherty, and R. C. Dougherty, *Digital Signal Processing*. Reston Publishing Company, 1984.

[13]
X. Li, S. Liu, M. W. Y. Lam, Z. Wu, C. Weng, and H. Meng, "Diverse and Expressive Speech Prosody Prediction with Denoising Diffusion Probabilistic Model." arXiv, Oct. 07, 2023. doi: 10.48550/arXiv.2305.16749.

[14]
Y. Gao, N. Morioka, Y. Zhang, and N. Chen, "E3 TTS: Easy End-to-End Diffusion-based Text to Speech." arXiv, Nov. 01, 2023. doi: 10.48550/arXiv.2311.00945.

[15]
T. Rajapakshe, R. Rana, S. Khalifa, B. Sisman, B. W. Schuller, and C. Busso, "emoDARTS: Joint Optimisation of CNN & Sequential Neural Network Architectures for Superior Speech Emotion Recognition." arXiv, Mar. 20, 2024. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/2403.14083

[16]
C.-B. Im, S.-H. Lee, S.-B. Kim, and S.-W. Lee, "EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 6317–6321. doi: 10.1109/ICASSP43922.2022.9747098.

[17]
D. Diatlova and V. Shutov, "EmoSpeech: Guiding FastSpeech2 Towards Emotional Text to Speech." arXiv, Jun. 28, 2023. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/2307.00024

[18]
S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional Speech Synthesis with Rich and Granularized Control," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7254–7258. doi: 10.1109/ICASSP40776.2020.9053732.

[19]
M. Schröder, "Emotional speech synthesis: a review," Sep. 2001, pp. 561–564. doi: 10.21437/Eurospeech.2001-150.

[20]
A. Erell and M. Weintraub, "Estimation using Log-spectral-distance criterion for Noise-robust speech recognition," May 1990, pp. 853–856 vol.2. doi: 10.1109/ICASSP.1990.115972.

[21]
N. Tits, K. E. Haddad, and T. Dutoit, "Exploring Transfer Learning for Low Resource Emotional TTS." arXiv, Jan. 14, 2019. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/1901.04276

[22]
Y. Ren *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech." arXiv, Aug. 07, 2022. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/2006.04558

[23]
Y. Lei, S. Yang, and L. Xie, "Fine-grained Emotion Strength Transfer, Control and Prediction for Emotional Speech Synthesis." arXiv, Nov. 17, 2020. doi: 10.48550/arXiv.2011.08477.

[24]
W. Wang, S. Xu, and B. Xu, "First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention," presented at the Proc. Interspeech 2016, 2016, pp. 2243–2247. doi: 10.21437/Interspeech.2016-134.

[25]
D. Ma, Z. Su, W. Wang, and Y. Lu, "FPETS : Fully Parallel End-to-End Text-to-Speech System." arXiv, Feb. 09, 2020. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/1812.05710

[26]
A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, Mar. 2020, doi: 10.1016/j.physd.2019.132306.

[27]
I. J. Goodfellow *et al.*, "Generative Adversarial Networks." arXiv, Jun. 10, 2014. doi: 10.48550/arXiv.1406.2661.

[28]
I. J. Goodfellow *et al.*, "Generative Adversarial Networks." arXiv, Jun. 10, 2014. doi: 10.48550/arXiv.1406.2661.

[29]
J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 8067–8077. Accessed: May 24, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/5c3b99e8f92532e5ad1556e53ceea00c-Abstract.html

[30]
D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions." arXiv, Jul. 10, 2018. doi: 10.48550/arXiv.1807.03039.

[31]
J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis." arXiv, Oct. 23, 2020. doi: 10.48550/arXiv.2010.05646.

[32]
"IEMOCAP- Home." Accessed: May 24, 2024. [Online]. Available: https://sail.usc.edu/iemocap/

[33]
X. An, F. K. Soong, and L. Xie, "Improving Performance of Seen and Unseen Speech Style Transfer in End-to-end Neural TTS." arXiv, Jun. 18, 2021. doi: 10.48550/arXiv.2106.10003.

[34]
D.-R. Liu, C.-Y. Yang, S.-L. Wu, and H.-Y. Lee, "Improving Unsupervised Style Transfer in end-to-end Speech Synthesis with end-to-end Speech Recognition," in *2018 IEEE Spoken*

Language Technology Workshop (SLT), Dec. 2018, pp. 640–647. doi: 10.1109/SLT.2018.8639672.

[35]

M. Edgington, "Investigating the limitations of concatenative synthesis," presented at the Proc. Eurospeech 1997, 1997, pp. 593–596. doi: 10.21437/Eurospeech.1997-217.

[36]

K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model architectures to extrapolate emotional expressions in DNN-based text-to-speech," *Speech Communication*, vol. 126, pp. 35–43, Feb. 2021, doi: 10.1016/j.specom.2020.11.004.

[37]

A. Bapna *et al.*, "mSLAM: Massively multilingual joint pre-training for speech and text." arXiv, Feb. 02, 2022. doi: 10.48550/arXiv.2202.01374.

[38]

F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions." arXiv, Apr. 30, 2016. doi: 10.48550/arXiv.1511.07122.

[39]

C. Lu, X. Wen, R. Liu, and X. Chen, "Multi-Speaker Emotional Speech Synthesis with Fine-Grained Prosody Modeling," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 5729–5733. doi: 10.1109/ICASSP39728.2021.9413398.

[40]

Z. Ju *et al.*, "NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models." arXiv, Apr. 23, 2024. doi: 10.48550/arXiv.2403.03100.

[41]

X. Tan *et al.*, "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality." arXiv, May 10, 2022. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/2205.04421

[42]

N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural Speech Synthesis with Transformer Network." arXiv, Jan. 30, 2019. doi: 10.48550/arXiv.1809.08895.

[43]

F. Bao *et al.*, "One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale." arXiv, May 30, 2023. doi: 10.48550/arXiv.2303.06555.

[44]

"openslr.org." Accessed: May 24, 2024. [Online]. Available: https://openslr.org/60/

[45]

"Papers with Code - Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers." Accessed: May 23, 2024. [Online]. Available: https://paperswithcode.com/paper/self-attention-a-better-building-block-for

[46]

C. M. Bishop, *Pattern recognition and machine learning*. in Information science and statistics. New York: Springer, 2006.

[47]

D. Kumawat and V. Jain, "POS Tagging Approaches: A Comparison," *International Journal of Computer Applications*, vol. 118, no. 6, pp. 32–38, May 2015.

[48]

Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable Text-to-Speech with Text Descriptions." arXiv, Nov. 22, 2022. doi: 10.48550/arXiv.2211.12171.

[49]

K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 972–980, May 2006, doi: 10.1109/TSA.2005.858051.

[50]

H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds1," *Speech Communication*, vol. 27, no. 3, pp. 187–207, Apr. 1999, doi: 10.1016/S0167-6393(98)00085-5.

[51]

Z. Chen, G. He, K. Zheng, X. Tan, and J. Zhu, "Schrodinger Bridges Beat Diffusion Models on Text-to-Speech Synthesis." arXiv, Dec. 06, 2023. doi: 10.48550/arXiv.2312.03491.

[52]

D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984, doi: 10.1109/TASSP.1984.1164317.

[53]

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, ISCA, Sep. 1999, pp. 2347–2350. doi: 10.21437/Eurospeech.1999-513.

[54]

H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1997, pp. 1303–1306 vol.2. doi: 10.1109/ICASSP.1997.596185.

[55]

K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013, doi: 10.1109/JPROC.2013.2251852.

[56]

"Speech Synthesis Markup Language (SSML) | Cloud Text-to-Speech API | Google Cloud." Accessed: May 23, 2024. [Online]. Available: https://cloud.google.com/text-to-speech/docs/ssml

[57]

"Speech Synthesis Markup Language (SSML) Version 1.1." Accessed: May 23, 2024. [Online]. Available: https://www.w3.org/TR/speech-synthesis11/

[58]

A. W. Black, H. Zen, and K. Tokuda, “Statistical Parametric Speech Synthesis”.

[59]

H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada: IEEE, May 2013, pp. 7962–7966. doi: 10.1109/ICASSP.2013.6639215.

[60]

Y. Wang *et al.*, “Tacotron: Towards End-to-End Speech Synthesis.” arXiv, Apr. 06, 2017. doi: 10.48550/arXiv.1703.10135.

[61]

H. Zen *et al.*, “The HMM-based Speech Synthesis System (HTS) Version 2.0,” 2007.

[62]

S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[63]

J. H. Ro, F. Stahlberg, K. Wu, and S. Kumar, “Transformer-based Models of Text Normalization for Speech Applications.” arXiv, Jan. 31, 2022. doi: 10.48550/arXiv.2202.00153.

[64]

Y. Fan, Y. Qian, F.-L. Xie, and F. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” Sep. 2014, pp. 1964–1968. doi: 10.21437/Interspeech.2014-443.

[65]

E. Song *et al.*, “TTS-by-TTS 2: Data-selective augmentation for neural speech synthesis using ranking support vector machine with variational autoencoder.” arXiv, Jun. 29, 2022. doi: 10.48550/arXiv.2206.14984.

[66]

H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 4470–4474. doi: 10.1109/ICASSP.2015.7178816.

[67]

N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, “Visualization and Interpretation of Latent Spaces for Controlling Expressive Speech Synthesis through Audio Analysis.” arXiv, Mar. 27, 2019. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/1903.11570

[68]

E. A. AlBadawy, A. Gibiansky, Q. He, J. Wu, M.-C. Chang, and S. Lyu, “VocBench: A Neural Vocoder Benchmark for Speech Synthesis.” arXiv, Dec. 06, 2021. Accessed: May 24, 2024. [Online]. Available: http://arxiv.org/abs/2112.03099

[69]

L. Rabiner and M. Sambur, “Voiced-unvoiced-silence detection using the Itakura LPC distance measure,” in *ICASSP ’77. IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 1977, pp. 323–326. doi: 10.1109/ICASSP.1977.1170330.

[70]

Y.-A. Chung *et al.*, “W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training.” arXiv, Sep. 13, 2021. doi: 10.48550/arXiv.2108.06209.

[71]

R. Prenger, R. Valle, and B. Catanzaro, “WaveGlow: A Flow-based Generative Network for Speech Synthesis.” arXiv, Oct. 30, 2018. doi: 10.48550/arXiv.1811.00002.

[72]

N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “WaveGrad: Estimating Gradients for Waveform Generation.” arXiv, Oct. 09, 2020. doi: 10.48550/arXiv.2009.00713.

[73]

A. van den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio.” arXiv, Sep. 19, 2016. Accessed: May 23, 2024. [Online]. Available: http://arxiv.org/abs/1609.03499

[74]

“WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications.” Accessed: May 24, 2024. [Online]. Available: https://www.jstage.jst.go.jp/article/transinf/E99.D/7/E99.D_2015EDP7457/_article