УДК 004.934.5:534.782
DOI: 10.18523/2617-3808.2025.8.28-37

D. Ivashchenko, O. Marchenko

MODERN APPROACHES TO CONTROLLABLE EMOTIONAL SPEECH SYNTHESIS

The generation of emotionally expressive and controllable speech is one of the most dynamic and technically demanding areas in the intersection of artificial intelligence, natural language processing, and speech synthesis. Recent progress in emotional text-to-speech (TTS) systems has enabled increasingly natural and emotionally nuanced voice generation, shifting from early concatenative methods to advanced neural models. This review provides a structured overview of the state of the art in controllable emotional TTS, highlighting key architectural paradigms. A special focus is placed on emotional control mechanisms, including discrete emotional tagging with categorical or dimensional labels, reference-based control which conditions synthesis on prosodic or stylistic exemplars, and prompt-based techniques that leverage the capabilities of large language models for flexible and intuitive emotional specification.

Despite substantial improvements in synthesis quality and emotional expressiveness, several critical challenges remain unresolved. These include the disentanglement of emotional, speaker, and prosodic features, the lack of standardized evaluation metrics for emotional clarity and naturalness, and the significant computational demands associated with training high-fidelity models. Furthermore, the scarcity of diverse and emotion-labeled speech data, especially for low-resource and morphologically rich languages, continues to limit the generalizability of current approaches. This review not only summarizes existing methods and their trade-offs but also outlines promising research directions, aiming to support the development of more robust, efficient, and emotionally expressive speech generation systems.

Keywords: deep learning, text-to-speech synthesis, natural language processing, speech emotion control, diffusion models.

Introduction

One of the most difficult but important tasks in developing natural language processing and artificial intelligence is generating speech with predefined characteristics and style. Generating natural, emotionally expressive speech has long been a central goal in artificial intelligence and natural language processing. Achieving this requires synthesizing speech that sounds humanlike and conveys emotional nuances, bridging the gap between human communication and machine-generated speech. Such advancements hold immense promise for applications ranging from assistive technologies and virtual assistants to entertainment and education. However, achieving this level of sophistication remains a formidable challenge, particularly in low-resource language contexts where labeled emotional datasets are limited.

Traditional text-to-speech (TTS) systems relied mainly on deterministic algorithms that frequently failed to produce dynamic, natural-sounding speech. Recent advances in deep learning have transformed TTS by allowing models to capture complex prosodic and emotive variations. While several systems have aimed to improve naturalness, few have investigated the explicit regulation of emotional expression in synthesized speech [34]. This feature is crucial for developing more engaging and context-appropriate human–computer interactions.

There are a number of basic obstacles that affect the capacity to produce emotionally expressive discourse. These include gathering enough emotion-labeled training data, effectively simulating the intricate link between text content and emotional expression, and providing intuitive controls for the emotional aspects of synthesized speech [5, 34]. These issues have been largely addressed by recent developments using innovative designs and training techniques.

This review examines the current state of emotional speech synthesis, focusing on recent advances in controllable and expressive TTS systems. We analyze the emotion control approaches that represent an important innovation in this domain. By examining its technical foundations, methodological approaches,

and evaluation results, we aim to provide a comprehensive overview of the current state of the art in emotional speech synthesis and identify promising directions for future research.

Text-to-Speech Synthesis Approaches

Non-Neural Approaches to Speech Synthesis

Text-to-speech conversion has historically progressed through several distinct paradigms. Early systems relied on concatenative synthesis, piecing together pre-recorded speech fragments to form complete utterances. While providing natural-sounding elements, these systems struggled with smooth transitions and had limited flexibility. The subsequent development of statistical parametric speech synthesis (SPSS), particularly Hidden Markov Model (HMM)-based approaches, offered greater control but often at the cost of naturalness [24].

Neural TTS Pipeline

The advent of deep learning has revolutionized TTS technology. Modern neural TTS systems can be categorized into several architectural approaches, each with distinct characteristics that affect their overall performance and application.

Mu, Yang, and Dong (2021) summarize TTS system construction approaches and methods. They articulate a modern speech generation pipeline as an end-to-end system, comprising three parts: a text-analysis front end, an acoustic model, and a vocoder [24]. The process begins with the text front end transforming the input text into a standardized format. Next, the acoustic model processes this standardized input into intermediate acoustic features, incorporating long-term structures from the speech. Common representations include spectrograms, vocoder features, and linguistic features. Lastly, the vocoder generates the final output by adding fine-grained signal details and converting the acoustic features into a time-domain waveform. Each system component has its unique architecture, which defines its application (Figure 1). Another category of methods follows a fully end-to-end philosophy, in which a single model performs all the necessary processing steps. It directly generates speech audio output from the input data, skipping the creation of intermediate features, such as mel-spectrograms, is skipped entirely [34].

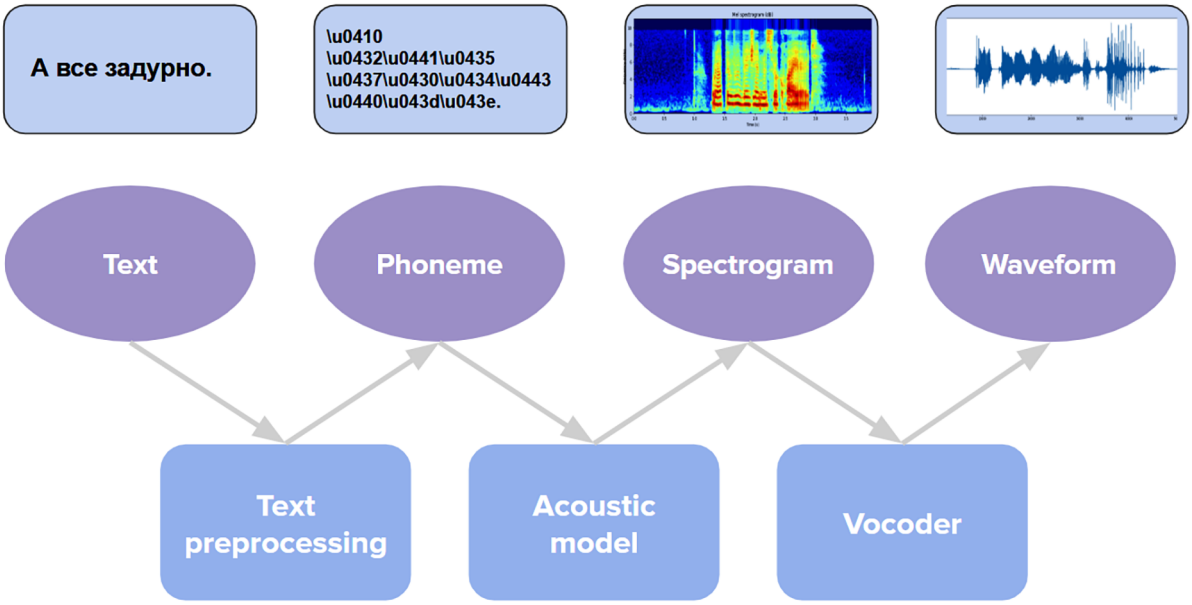

Figure 1. Mainstream TTS pipeline

There are a large volume of published studies describing a wide variety of approaches to neural speech generation. The systems can be conventionally divided into distinct segments based on the generation technology employed in a particular system. This review primarily considers architectures that are widely used for expressive speech synthesis.

Controllable TTS Systems Architectures

Autoregressive and Non-Autoregressive Generation Models

Autoregressive generation models have transformed speech synthesis by predicting speech representations sequentially, with each output based on all previously generated tokens, allowing for the capture of intricate temporal relationships inherent in human speech [24]. These models include a variety of acoustic architectures, such as transformer-based systems, convolutional neural networks (CNNs), recurrent neural networks (RNNs), flow-based models, and diffusion-based approaches, each with unique advantages for modeling speech characteristics and improving generation quality [3]. Autoregressive speech synthesis has made significant progress in producing genuine, expressive speech that closely resembles human voice patterns by learning directly from data rather than relying on substantial manual feature engineering.

Non-autoregressive speech synthesis models address the computational inefficiency of sequential generation by allowing parallel prediction of acoustic features, which reduces inference time from several seconds to real-time or faster production rates [24]. These architectures also encompass a variety of generative approaches, including feed-forward transformer networks, flow-based models, generative adversarial networks, variational autoencoders, and diffusion-based systems, all of which use parallel computation to eliminate sequential dependencies inherent in autoregressive methods [34]. While avoiding error propagation difficulties provides significant speedups and greater stability, non-autoregressive algorithms frequently require explicit duration prediction or external alignment mechanisms to handle the one-to-many mapping challenge between text and speech sequences [15].

Main Neural Architecture Types

Tacotron [33] presented the first acoustic model based on deep learning and remains used in various systems. A one-dimensional convolution- and bidirectional gated recurrent unit-based encoder, an attention-driven decoder, and a Griffin-Lim vocoder are the main parts of the proposed autoregressive architecture, which is a sequence-to-sequence (seq2seq) model. It accepts characters as input and generates spectrograms, which are subsequently transformed into waveforms by the vocoder.

Later, Ren et al. presented FastSpeech 2 (2020) [12], an enhanced iteration of the FastSpeech model. It extends parallel feed-forward transformer with length prediction to rectify the constraints of its predecessor, attaining accelerated training speed and providing control over speech variation information (e.g., pitch, energy, and more accurate duration) as conditional inputs with fully end-to-end text-to-waveform synthesis.

Another type of speech synthesizer models the complex speech distribution as a repetitive composition of simple distributions, which is called a generative flow (Glow) [24]. Previously, the concept was mainly used in vocoders of neural TTS, but now it is also applied to acoustic models [3]. Valle et al. introduced Flowtron, an autoregressive text-to-speech synthesis model based on autoregressive flow techniques [13]. This novel generative network achieves high-quality mel-spectrograms and enables the manipulation of speech variation and style transfer. Moreover, it possesses the capability to control various aspects of speech, including pitch, tone, speech rate, cadence, and accent.

Diffusion neural networks are probabilistic generative models, which operate by learning to reverse a gradual noise corruption process. This enables speech generation through iterative denoising of random noise conditioned on text input and control parameters [2]. The stochastic nature of diffusion processes allows for diverse synthesis outputs while maintaining stable training dynamics, effectively addressing the main issues that have historically challenged other controllable generative approaches in speech synthesis. StyleTTS 2 [32] is a non-autoregressive TTS model with differentiable duration modeling, leveraging style diffusion and adversarial training with large speech language models (SLMs) to achieve human-level synthesis quality. The system is capable of synthesizing contextually appropriate styles directly from reference speech in a zero-shot scenario.

Generative Adversarial Networks (GANs) have become an important technology in controllable TTS systems [34]. GANs consist of a generator that synthesizes speech from text and a discriminator that evaluates its realism, trained adversarially to produce high-fidelity outputs. GANs are widely used in vocoders, as well as acoustic models for end-to-end systems [3, 24, 34]. Among vocoders, HiFi-GAN [19] is a state-of-the-art GAN-based approach that achieves both high-fidelity speech synthesis and exceptional computational efficiency, which implies the diversity of applications in end-to-end speech systems. It employs in-

novative multi-scale and multi-period discriminators, combined with a generator incorporating multi-receptive field fusion modules, which effectively capture diverse temporal patterns in speech through dilated convolutions with different dilation rates and kernel sizes.

Classification of Emotional TTS Models

Emotional Tagging Control

Emotional tagging represents the most straightforward approach to expressiveness control, using explicit labels or encodings to condition synthesis models. This methodology has evolved from simple one-hot encodings to sophisticated multi-dimensional representations.

Older systems allow users to select from a predefined set of discrete emotion labels, such as happy, sad, angry, or neutral [5]. The model is usually trained on datasets where speech samples are annotated with these labels, such as ESD, RAVDESS, and IEMOCAP, enabling it to generate speech with the corresponding emotional tone.

The label-based method can be implemented in multiple modes. For instance, MsEmoTTS [23] employs a multi-scale emotional speech synthesis framework that can be conditioned with one-hot encoded vectors representing discrete emotions such as happiness, anger, sadness, surprise, fear, and disgust. Inoue et al. introduce a hierarchical emotion distribution (HED) model for continuous emotion intensity control across phonemes, words, and utterances [16]. Practically, the expressiveness is controlled via the vector of emotional intensities from 0 to 1. In contrast, StyleTagging-TTS (ST-TTS) [11] and Emo-DPO [7] systems represent an approach that utilizes a defined set of style tags written in natural language to control emotional expression, modeling the relationship between linguistic embedding and speaking emotion domain with a pre-trained language model.

A significant drawback of this approach is the limitations in emotional intensity control. Systems, such as MsEmoTTS, have implicit predictors, but no explicit control over that characteristic is provided. Other challenges include customization issues due to fixed categories set and the need for large labeled datasets for training [5].

A different method for emotion encoding refers to representing it as a vector of so-called basic (or fundamental) emotions. Zhou et al. [30] present a study on modeling and synthesizing mixed emotions in speech synthesis. The paper extensively references Plutchik's emotion wheel theory, stating eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy, and arranges them in a framework where all the emotional styles can be derived from those [26]. The framework allows users to manually control emotion rendering by defining an emotion attribute vector with specific percentages for each primary emotion, successfully synthesizing complex emotional states like delight (surprise + happy), outrage (surprise + angry), and disappointment (surprise + sad) [30]. Similarly, for the EmoMix model [8], the authors considered excitement (happy + surprise) and disappointment (sad + surprise) within the intensity range for evaluation.

The PAD (Pleasure-Arousal-Dominance) model, also known as the VAD (Valence-Arousal-Dominance) model, is a three-dimensional psychological framework developed by Albert Mehrabian and James A. Russell to describe and measure emotional states using numerical dimensions [4]. The model represents emotions through valence (the pleasantness or unpleasantness of an emotion), arousal (the intensity or energy level associated with the emotion), and dominance (the degree of control or power one feels in the emotional state).

This dimensional approach has been widely adopted in emotional speech synthesis systems because it provides more nuanced emotional rendering than discrete categorical emotion models [10]. Sivaprasad et al. [29] extended the FastSpeech2 TTS architecture, introducing a prosody control block that conditions phoneme-level pitch, energy, and duration on continuous arousal-valence values, enabling fine-grained, interpretable emotional prosody control. EmoSphere-TTS [10] synthesizes expressive speech by embedding valence-arousal-dominance representation into a spherical vector via transformation. This approach enables manual control of emotional style and intensity by assigning appropriate angles and lengths to the spherical vector. UDDETTS framework [20] introduces a neural codec language model that seamlessly integrates traditional discrete emotion labels with the continuous VAD model, enabling controllable, expressive TTS through joint optimization in a unified emotional space.

The main drawback of continuous vector emotional representations is the scarcity of annotated training data, as dimensional emotion labels (e.g., arousal, valence, dominance) are costly and subjective to obtain

at scale. Moreover, models that leverage these representations often require complex architectures and careful tuning to prevent overfitting and ensure interpretable, controllable outputs.

Reference-based Control

Reference-based methods extract style information from example utterances, enabling control without explicit annotation. The reference audio provides a direct example of the emotional tone, which the system mimics. Reference-based systems use architectures where a reference encoder extracts emotional style embeddings from the input sample. This paradigm supports both voice conversion and style transfer applications.

Mellotron [22] is a multispeaker voice synthesis model based on Tacotron 2 [6], which can make a voice emote without emotive training data by explicitly conditioning on rhythm and continuous pitch contours from an audio signal. It pioneered reference-based emotional control by incorporating global style tokens (GST) [31] as learned latent variables alongside reference-based speech conditioning, allowing it to transfer text, rhythm, and pitch characteristics from source audio to target speakers while maintaining fine-grained control over expressive speech characteristics. Building upon such a reference-based approach, GenerSpeech [14] proposes a generalizable text-to-speech model that decomposes speech variation into style-agnostic and style-specific parts through a multi-level style adaptor and Mix-Style Layer Normalization, enabling robust zero-shot style transfer for out-of-domain custom voices without requiring adaptation data.

StyleTTS [21] extends the approach by using style encoders that extract emotion-relevant features while disentangling them from speaker identity and linguistic content. The system employs adversarial training to ensure the sound quality of the reconstructed mel-spectrogram. SC VALL-E [18] adopts the VALL-E [25] neural codec language model, consideri speech synthesis as a conditional language modeling task, and uses reference audio for speaker cloning and style transfer, including speaking rate, pitch, voice intensity, and emotional styles. EmoSphere++ [9] extends the EmoSphere-TTS [10] model and introduces an emotion-adaptive spherical vector that extracts it directly from reference speech samples, using a multi-level style encoder to capture both high-level emotional categories and low-level nuanced expressions for zero-shot emotion transfer.

Reference-based TTS systems face significant challenges in their dependency on high-quality reference audio and struggle with generalizing to unseen emotions, speakers, or prosodic styles not represented in the training data, while also requiring effective disentanglement of emotional content from other style features like speaker identity and linguistic prosody. Additionally, these approaches encounter practical limitations, including computational overhead from processing reference audio during inference and the scarcity of fine-grained explicit emotion control with this method.

Prompt-based Control

With the rapid development of large language models (LLMs) in recent years, prompt-based approaches represent the frontier of controllable emotional TTS, using large language models to interpret and execute complex synthesis instructions. These systems go beyond simple emotional labels to understand contextual and situational factors and allow users to specify emotions using natural language descriptions or instructions, e.g., "speak happily" or "whisper fearfully." This approach is user-friendly, leveraging natural language processing (NLP) to interpret prompts.

PromptTTS [28] utilizes a BERT-based style encoder to extract semantic representations from text prompts, conditioning a FastSpeech 2-based non-autoregressive acoustic model for mel-spectrogram generation, paired with a HiFi-GAN vocoder for waveform synthesis. Its successor, PromptTTS 2 [27], integrates LLMs for enhanced prompt understanding and a diffusion-based variation network to model diverse emotional styles, improving expressiveness and voice variability. InstructTTS [17] employs a VQ-VAE to discretize speech into latent codes, using a diffusion transformer to align text prompts with these codes, enabling fine-grained emotional control via a BERT-like approach, and a RoBERTa encoder for instruction processing. CosyVoice [6] combines an LLM to enable fine-grained freestyle natural language emotion control, using Qwen2.5-0.5B as its backbone with supervised semantic tokens, feeding into a non-autoregressive transformer decoder for multilingual zero-shot synthesis, with a style encoder for prompt-driven emotion control.

These systems demonstrate the field's convergence toward transformer-based architectures for interpreting emotional descriptions. However, challenges around prompt ambiguity and computational complexity

persist. Yet, translating natural language emotion descriptions to consistent acoustic realizations remains challenging due to the inherent subjectivity in emotional expressions and data availability for a prompt-to-style alignment system.

Quality Evaluation

Evaluating emotional TTS systems requires assessing both synthesis quality and emotional appropriateness, presenting unique challenges compared to neutral TTS evaluation.

Objective Metrics

Mel Cepstral Distortion (MCD) quantifies the spectral distance between synthesized and reference emotional speech [34]. Better speech synthesis quality is shown by a lower MCD value, which also indicates a higher resemblance between the reference and synthetic speech. Good quality is often indicated by an MCD value less than 4; however, substantial distortion may be indicated by values greater than 6. It can be calculated using the following formula:

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{i=1}^{M}\left(c_i^{gen} - c_i^{ref}\right)^2}$$

where c_i^{gen} and c_i^{ref} are the i-th mel cepstral coefficient (MCC) of generated and reference speech, respectively, and M is the total number of MCCs. While widely used and applicable in reference-based systems, MCD may not capture perceptually important emotional characteristics.

As pitch level is vastly impacted by emotional state, Gross Pitch Error (GPE) is considered for expressive TTS [5]. It determines the proportion of voiced frames with a pitch deviation of greater than a specific threshold (often 20%) [1]. Emotion and style classifiers and speech emotion recognition (SER) models are widely used to measure classification accuracy, reflecting the efficiency of the proposed model in generating emotional speech [5]. It is worth mentioning that this evaluation relies on appropriate SER method selection.

Cosine similarity between emotional embeddings quantifies how similar the synthesized and reference speech is in terms of emotional expression [34]. It can be used to evaluate emotion-controllable TTS methods, where higher values indicate better emotional similarity. Emotional embeddings can be extracted using pre-trained emotion recognition models, such as x-vector-based systems trained on emotional speech datasets, to quantify how well the synthesized speech matches the intended emotional expression of the reference audio.

The Word Error Rate (WER) [34] is used to ensure intelligibility. By calculating the amount of transcription errors, it measures the difference between the reference transcript and the recognized transcript. WER is calculated as follows:

$$WER = \frac{W + M + E}{N}$$

where W is the number of wrong words in place of the correct word, M is the number of missed words, E is the number of extra words added, and N is the total number of words in the reference transcript. However, this metric is highly dependent on the choice of transcription method.

Subjective metrics

The most widely used subjective statistic is the Mean Opinion Score (MOS) [34]. On a scale of 1 to 5, listeners judge many characteristics of synthesized speech, including naturalness and expressiveness [10], and the result is the average of these scores. Higher ratings denote higher quality. Although MOS is costly for extensive assessments, it successfully captures the human perspective [34]. Two TTS audio samples are

compared for relative quality differences using the Comparison Mean Opinion Score (CMOS) [5]. After hearing paired samples, participants score their preference on a scale, which usually includes both negative and positive values (e.g., from -3 to 3) [34], and then it is averaged as in MOS.

Various researchers have employed these metrics in various ways to evaluate expressivity-related characteristics. Specifically, the metrics have been tested on audio samples that encompass diverse emotional expressions, distinct speaking mannerisms, and different degrees of intensity [5]. Additionally, these samples span multiple speech synthesis contexts, including scenarios involving parallel versus non-parallel style conversion, as well as familiar versus unfamiliar speaking styles and voices.

Discussion

Emotional TTS systems pose considerable technological problems in terms of feature disentanglement and representation, making it difficult to distinguish emotional aspects from other speech features such as speaker identification or linguistic content. This can, however, be handled using specialized disentanglement architectures or adversarial training strategies that expressly require the separation of emotional and speaker-specific information. Development of interpretable latent space manipulation techniques and controllable generation methods using semantic emotion embeddings could potentially improve fine-grained control and interpretability.

Evaluation metrics lack objectivity and standardization, making it difficult to compare different approaches or accurately measure progress across studies, an issue which cannot be fully eliminated. The significant gap in generating computationally efficient models can be bridged using knowledge distillation approaches, in which lightweight student models learn from bigger teacher networks, or architectural improvements such as separable convolutions and parameter sharing algorithms.

Low-resource languages face even greater obstacles due to minimal emotional speech data availability, though cross-lingual transfer learning and multilingual pre-training approaches offer promising solutions to explore. Given the daunting challenge of building comprehensive emotional speech datasets spanning diverse emotions, styles, speakers, and languages, data augmentation and knowledge transfer techniques become essential for addressing data scarcity in expressive speech synthesis.

These challenges collectively limit practical deployment, but the convergence of these solution strategies creates opportunities for researchers to make meaningful contributions that advance multiple aspects of emotional TTS simultaneously.

Conclusion

The swift progressions in emotional text-to-speech (TTS) synthesis have transitioned from rudimentary concatenative methodologies to sophisticated neural architectures capable of generating emotionally resonant and high fidelity speech. This evolution has been significantly propelled by deep learning, which enables intricate management of emotional tone and stylistic variation.

In contemporary systems, transformer-based and diffusion-based architectures are increasingly prevalent. Transformers facilitate adjustable synthesis and parallel generation, while diffusion models employ iterative refinement to enhance quality. The previously utilized fixed emotive tags have been refined by reference-based and prompt-driven control mechanisms, and the incorporation of LLMs has created novel strategies for adaptable, and user-centric interaction.

Notwithstanding these advancements, several challenges persist. It remains an obstacle to disentangling emotion from speaker identification and content. The acquisition of high-quality emotional data continues to be a daunting endeavor, particularly within low-resource languages, and the lack of standardized evaluation leads to inconsistencies in comparative analyses.

Список літератури

1. A comparative study of pitch extraction algorithms on a large variety of singing sounds / Onur Babacan [et al.] // ICASSP 2013 - proceedings. – Vancouver, Canada, 2013. — Pp. 1–5.
2. A survey on audio diffusion models: text to speech synthesis and enhancement in generative AI [Electronic resource] / Chenshuang Zhang [et al.]. — Mode of access: arXiv.2303.13336.
3. A survey on neural speech synthesis [Electronic resource] / Xu Tan [et al.]. — Mode of access: arXiv.2106.15561.
4. Bales R. F. Social interaction systems: theory and measurement / Robert Freed Bales. — New Brunswick, NJ : Transaction Publishers, 2017.

5. Barakat H. Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources / Huda Barakat, Oytun Turk, Cenk Demiroglu // EURASIP journal on audio, speech, and music processing. — 2024. — Vol. 2024, no. 1. — P. 11.
6. CosyVoice: a scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens [Electronic resource] / Zhihao Du [et al.]. — Mode of access: arXiv.2407.05407.
7. Emo-DPO: controllable emotional speech synthesis through direct preference optimization [Electronic resource] / Xiaoxue Gao [et al.]. — Mode of access: arXiv.2409.10157.
8. EmoMix: emotion mixing via diffusion models for emotional speech synthesis / Haobin Tang [et al.] // Interspeech 2023. — [S. l.], 2023. — Pp. 12–16.
9. EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech via Emotion-Adaptive Spherical Vector / Deok-Hyeon Cho [et al.] // IEEE transactions on affective computing. — 2025. — Pp. 1–16.
10. EmoSphere-TTS: emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech / Deok-Hyeon Cho [et al.] // Interspeech 2024. — [S. l.], 2024. — Pp. 1810–1814.
11. Expressive text-to-speech using style tag / Minchan Kim [et al.] // Interspeech 2021. — [S. l.], 2021. — Pp. 4663–4667.
12. FastSpeech 2: fast and high-quality end-to-end text to speech [Electronic resource] / Yi Ren [et al.]. — Mode of access: arXiv.2006.04558.
13. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis [Electronic resource] / Rafael Valle [et al.] // International conference on learning representations. — [S. l.], 2021. — Mode of access: https://openreview.net/forum?id=Ig53hpHxS4.
14. GenerSpeech: towards style transfer for generalizable out-of-domain text-to-speech / Rongjie Huang [et al.] // Proceedings of the 36th international conference on neural information processing systems. — Red Hook, NY, USA, 2022.
15. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search / Jaehyeon Kim [et al.] // Advances in neural information processing systems. — [S. l.], 2020. — Pp. 8067–8077.
16. Hierarchical emotion prediction and control in text-to-speech synthesis / Sho Inoue [et al.] // ICASSP 2024 - 2024 IEEE international conference on acoustics, speech and signal processing (ICASSP). — [S. l.], 2024. — Pp. 10601–10605.
17. InstructTTS: modelling expressive TTS in discrete latent space with natural language style prompt / Dongchao Yang [et al.] // IEEE/ACM transactions on audio, speech, and language processing. — 2024. — Vol. 32. — Pp. 2913–2925.
18. Kim D. SC VALL-E: Style-Controllable Zero-Shot Text to Speech Synthesizer [Electronic resource] / Daegyeom Kim, Seongho Hong, Yong-Hoon Choi. — Mode of access: arXiv.2307.10550.
19. Kong J. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis / Jungil Kong, Jaehyeon Kim, Jaekyoung Bae // Advances in neural information processing systems. — [S. l.], 2020. — Pp. 17022–17033.
20. Liu J. UDDETTS: unifying discrete and dimensional emotions for controllable emotional text-to-speech [Electronic resource] / Jiaxuan Liu, Zhenhua Ling. — Mode of access: arXiv.2505.10599.
21. Li Y. A. StyleTTS: a style-based generative model for natural and diverse text-to-speech synthesis / Yinghao Aaron Li, Cong Han, Nima Mesgarani // IEEE journal of selected topics in signal processing. — 2025. — Vol. 19, no. 1. — Pp. 283–296.
22. Mellotron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens / Rafael Valle [et al.] // ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). — [S. l.], 2020. — Pp. 6189–6193.
23. MsEmoTTS: multi-scale emotion transfer, prediction, and control for emotional speech synthesis / Yi Lei [et al.] // IEEE/ACM trans. audio, speech and lang. proc. — 2022. — Vol. 30. — Pp. 853–864.
24. Mu Z. Review of end-to-end speech synthesis technology based on deep learning [Electronic resource] / Zhaoxi Mu, Xinyu Yang, Yizhuo Dong. — Mode of access: arXiv.2104.09995.
25. Neural codec language models are zero-shot text to speech synthesizers [Electronic resource] / Chengyi Wang [et al.]. — Mode of access: arXiv.2301.02111.
26. Plutchik R. Theories of emotion / Robert Plutchik, Henry Kellerman. — Burlington : Elsevier Science, 1980.
27. PromptTTS 2: describing and generating voices with text prompt / Yichong Leng [et al.] // The twelfth international conference on learning representations. — [S. l.], 2024.
28. Promptttts: controllable text-to-speech with text descriptions / Zhifang Guo [et al.] // ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). — [S. l.], 2023. — Pp. 1–5.
29. Sivaprasad S. Emotional prosody control for speech generation / Sarath Sivaprasad, Saiteja Kosgi, Vineet Gandhi // Interspeech 2021. — [S. l.], 2021. — Pp. 4653–4657.
30. Speech synthesis with mixed emotions / Kun Zhou [et al.] // IEEE transactions on affective computing. — 2023. — Vol. 14, no. 4. — Pp. 3120–3134.
31. Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis / Yuxuan Wang [et al.] // Proceedings of the 35th international conference on machine learning. — [S. l.], 2018. — Pp. 5180–5189.
32. StyleTTS 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models / Yinghao Aaron Li [et al.] // Advances in neural information processing systems / ed. by A. Oh [et al.]. — [S. l.], 2023. — Pp. 19594–19621.
33. Tacotron: towards end-to-end speech synthesis / Yuxuan Wang [et al.] // Interspeech 2017. — [S. l.], 2017. — Pp. 4006–4010.
34. Towards controllable speech synthesis in the era of large language models: a survey [Electronic resource] / Tianxin Xie [et al.]. — Mode of access: arXiv.2412.06602.

References

Babacan, O., Drugman, T., d'Alessandro, N., Henrich Bernardoni, N., & Dutoit, T. (2013). A comparative study of pitch extraction algorithms on a large variety of singing sounds. *ICASSP 2013 - Proceedings*, 1–5. https://hal.science/hal-00923967.

Bales, R. F. (2017). *Social interaction systems: Theory and measurement*. Transaction Publishers. https://doi.org/10.4324/9781315129563.

Barakat, H., Turk, O., & Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, *2024* (1), 11. https://doi.org/10.1186/s13636-024-00329-7.

Cho, D.-H., Oh, H.-S., Kim, S.-B., Lee, S.-H., & Lee, S.-W. (2024). EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech. *Interspeech 2024*, 1810–1814. https://doi.org/10.21437/Interspeech.2024-398.

Cho, D.-H., Oh, H.-S., Kim, S.-B., & Lee, S.-W. (2025). EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech via Emotion-Adaptive Spherical Vector. *IEEE Transactions on Affective Computing*, 1–16. https://doi.org/10.1109/TAFFC.2025.3561267.

Du, Z., Chen, Q., Zhang, S., Hu, K., Lu, H., Yang, Y., Hu, H., Zheng, S., Gu, Y., Ma, Z., Gao, Z., & Yan, Z. (2024). *CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens* (No. arXiv:2407.05407). arXiv. https://doi.org/10.48550/arXiv.2407.05407.

Gao, X., Zhang, C., Chen, Y., Zhang, H., & Chen, N. F. (2024). *Emo-DPO: Controllable Emotional Speech Synthesis through Direct Preference Optimization* (No. arXiv:2409.10157). arXiv. https://doi.org/10.48550/arXiv.2409.10157.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., & Tan, X. (2023). Prompttts: Controllable Text-To-Speech With Text Descriptions. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096285.

Huang, R., Ren, Y., Liu, J., Cui, C., & Zhao, Z. (2022). GenerSpeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Inoue, S., Zhou, K., Wang, S., & Li, H. (2024). Hierarchical Emotion Prediction and Control in Text-to-Speech Synthesis. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10601–10605.

Kim, D., Hong, S., & Choi, Y.-H. (2023). *SC VALL-E: Style-Controllable Zero-Shot Text to Speech Synthesizer* (No. arXiv:2307.10550). arXiv. https://doi.org/10.48550/arXiv.2307.10550.

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Advances in Neural Information Processing Systems*, *33*, 8067–8077.

Kim, M., Cheon, S. J., Choi, B. J., Kim, J. J., & Kim, N. S. (2021). Expressive Text-to-Speech Using Style Tag. *Interspeech 2021*, 4663–4667. https://doi.org/10.21437/Interspeech.2021-465.

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, *33*, 17022–17033.

Lei, Y., Yang, S., Wang, X., & Xie, L. (2022). MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, *30*, 853–864. https://doi.org/10.1109/TASLP.2022.3145293.

Leng, Y., Guo, Z., Shen, K., Tan, X., Ju, Z., Liu, Y., Liu, Y., Yang, D., Zhang, L., Song, K., He, L., Li, X.-Y., Zhao, S., Qin, T., & Bian, J. (2024). PromptTTS 2: Describing and Generating Voices with Text Prompt. *The Twelfth International Conference on Learning Representations*.

Li, Y. A., Han, C., & Mesgarani, N. (2025). StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis. *IEEE Journal of Selected Topics in Signal Processing*, *19* (1), 283–296. https://doi.org/10.1109/JSTSP.2025.3530171.

Li, Y. A., Han, C., Raghavan, V., Mischler, G., & Mesgarani, N. (2023). StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. B A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 19594–19621). Curran Associates, Inc.

Liu, J., & Ling, Z. (2025). *UDDETTS: Unifying Discrete and Dimensional Emotions for Controllable Emotional Text-to-Speech* (No. arXiv:2505.10599). arXiv. https://doi.org/10.48550/arXiv.2505.10599.

Mu, Z., Yang, X., & Dong, Y. (2021). *Review of end-to-end speech synthesis technology based on deep learning* (No. arXiv:2104.09995). arXiv. https://doi.org/10.48550/arXiv.2104.09995.

Plutchik, R., & Kellerman, H. (1980). *Theories of Emotion*. Elsevier Science.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *International Conference on Learning Representations*. https://openreview.net/forum?id=piLPYqxtWuA.

Sivaprasad, S., Kosgi, S., & Gandhi, V. (2021). Emotional Prosody Control for Speech Generation. *Interspeech 2021*, 4653–4657. https://doi.org/10.21437/Interspeech.2021-307.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). *A Survey on Neural Speech Synthesis* (No. arXiv:2106.15561). arXiv. https://doi.org/10.48550/arXiv.2106.15561.

Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023). EmoMix: Emotion Mixing via Diffusion Models for Emotional Speech Synthesis. *INTERSPEECH 2023*, 12–16. https://doi.org/10.21437/Interspeech.2023-1317.

Valle, R., Li, J., Prenger, R., & Catanzaro, B. (2020). Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6189–6193. https://doi.org/10.1109/ICASSP40776.2020.9054556.

Valle, R., Shih, K. J., Prenger, R., & Catanzaro, B. (2021). Flowtron: An Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis. *International Conference on Learning Representations*. https://openreview.net/forum?id=Ig53hpHxS4.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., & Wei, F. (2023). *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers* (No. arXiv:2301.02111). arXiv. https://doi.org/10.48550/arXiv.2301.02111.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., & Saurous, R. A. (2018). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *Proceedings of the 35th International Conference on Machine Learning*, 5180–5189. https://proceedings.mlr.press/v80/wang18h.html.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. *Interspeech 2017*, 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452.

Xie, T., Rong, Y., Zhang, P., Wang, W., & Liu, L. (2025). *Towards Controllable Speech Synthesis in the Era of Large Language Models: A Survey* (No. arXiv:2412.06602; arXiv. https://doi.org/10.48550/arXiv.2412.06602.

Yang, D., Liu, S., Huang, R., Weng, C., & Meng, H. (2024). InstructTTS: Modelling Expressive TTS in Discrete Latent Space With Natural Language Style Prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 2913–2925. https://doi.org/10.1109/TASLP.2024.3402088.

Zhang, C., Zhang, C., Zheng, S., Zhang, M., Qamar, M., Bae, S.-H., & Kweon, I. S. (2023). *A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI* (No. arXiv:2303.13336). arXiv. https://doi.org/10.48550/arXiv.2303.13336.

Zhou, K., Sisman, B., Rana, R., Schuller, B. W., & Li, H. (2023). Speech Synthesis With Mixed Emotions. *IEEE Transactions on Affective Computing*, *14* (4), 3120–3134. https://doi.org/10.1109/TAFFC.2022.3233324.

Іващенко Д. С., Марченко О. О.

СУЧАСНІ ПІДХОДИ ДО КОНТРОЛЬОВАНОГО СИНТЕЗУ ЕМОЦІЙНОГО МОВЛЕННЯ

У статті представлено комплексний огляд сучасних технологій керованих систем для емоційного синтезу мовлення. Проаналізовано еволюцію нейронних архітектур, систематизовано підходи за технологіями та методами емоційного контролю. Визначено ключові виклики галузі, що охоплюють відокремлення мовленнєвих ознак та дефіцит даних для мов з обмеженими ресурсами. Окреслено перспективні напрями розвитку систем емоційно контрольованого синтезу мовлення.

Ключові слова: глибоке навчання, синтез мовлення з тексту, обробка природної мови, емоційний контроль мовлення, дифузійні моделі.

Матеріал надійшов 25.06.2025