DOI: https://doi.org/10.26565/2304-6201-2024-61-05

УДК (UDC) 519.688:004.934

Ivaniuk Andrii
PhD Student
National University of "Kyiv-Mohyla Academy", Faculty of Computer Sciences, 2 Skovorody st., Kyiv, Ukraine, 04655
e-mail: a.ivaniuk@ukma.edu.ua
https://orcid.org/0000-0002-4189-3787

Latent diffusion model for speech signal processing

Topicality. The development of generative models for audio synthesis, including text-to-speech (TTS), text-to-music, and text-to-audio applications, largely depends on their ability to handle complex and varied input data. This paper centers on latent diffusion modeling, a versatile approach that leverages stochastic processes to generate high-quality audio outputs.
Key goals. This study aims to evaluate the efficacy of latent diffusion modeling for TTS synthesis on the EmoV-DB dataset, which features multi-speaker recordings across five emotional states, and to contrast it with other generative techniques.
Research methods. We applied latent diffusion modeling to TTS synthesis specifically and evaluated its performance using metrics that assess intelligibility, speaker similarity, and emotion preservation in the generated audio signal.
Results. The study reveals that while the proposed model demonstrates decent efficiency in maintaining speaker characteristics, it is outperformed by the discrete autoregressive model: xTTS v2 in all assessed metrics. Notably, the researched model exhibits deficiencies in emotional classification accuracy, suggesting potential misalignment between the emotional intents encoded by the embeddings and those expressed in the speech output.
Conclusions. The findings suggest that further refinement of the encoder's ability to process and integrate emotional data could enhance the performance of the latent diffusion model. Future research should focus on optimizing the balance between speaker and emotion characteristics in TTS models to achieve a more holistic and effective synthesis of human-like speech.

Keywords: *audio modeling, artificial neural networks, speech synthesis.*

Як цитувати: Ivaniuk A. Latent diffusion model for speech signal processing. *Вісник Харківського національного університету імені В.Н. Каразіна, сер. «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»*. 2024. вип. 61. С.43-52. https://doi.org/10.26565/2304-6201-2024-61-05

How to quote: Ivaniuk A., "Latent diffusion model for speech signal processing." *Bulletin of V.N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 61, pp. 43-52, 2024. https://doi.org/10.26565/2304-6201-2024-61-05

1. Introduction

The pursuit of sophisticated generative models has invariably involved the integration of robust frameworks that underpin the generation process. At the heart of our study is the development of a model designed to cater specifically to the complexities inherent in generating realistic and nuanced audio outputs. This model is not only pivotal for understanding the theoretical underpinnings of audio synthesis but also serves as the backbone for practical applications in various audio generation tasks, including Text-to-Speech (TTS) systems.

TTS technology, which converts text into spoken voice output, has seen significant advancements through the adoption of deep learning models that improve naturalness and intelligibility. To enhance our model's capabilities within the TTS domain, we align our objectives with those of existing implementations that leverage similar neural architectures. Notably, the implementation of the Audio Latent Diffusion Model 2 (Audio LDM2) provides a basis for its innovative approach to audio synthesis. This model, known for its effectiveness in handling high-dimensional audio data through a diffusion-based process, aligns closely with our goals.

The existence of such a model as Audio LDM2 offers an opportunity to not only refine our approach by tuning our model based on this established framework but also to rigorously compare its performance against current competitive models like xTTS v2. This comparative analysis aims to highlight the limitations and potential our approach brings to the TTS research field, potentially setting new benchmarks for audio quality.

2. Related Work

Text-to-speech (TTS) synthesis has seen significant advancements due to the adoption of deep learning techniques, which have greatly improved the naturalness and expressiveness of synthesized speech. This section reviews several key methodologies in TTS that share, particularly focusing on models that integrate advanced neural network architectures and embeddings to enhance speech quality and emotional expressivity.

Tacotron models. One of the foundational models in modern TTS is Tacotron, which uses a sequence-to-sequence framework with attention to convert text directly into speech [1]. This model laid the groundwork for further developments in end-to-end speech synthesis. Following Tacotron, Tacotron 2 integrated WaveNet, a deep generative model of raw audio waveforms, to improve the naturalness of the speech output [2].

Embedding-Based Models. Significant similarities can be drawn with models that utilize embeddings to capture speaker characteristics and emotional states. For instance, VoiceLoop uses a phoneme-level language model to generate speech from text while preserving the speaker's voice by incorporating speaker-specific embeddings [3]. Similarly, Emotional TTS systems often rely on emotion embeddings to modulate the speech output to convey different emotional tones [4].

Contrastive Learning. The use of contrastive learning, as seen in Contrastive Language-Audio Pretraining (CLAP) [5], is a relatively new trend. Models like HuBERT and WavLM have shown that pretraining audio models on largescale unlabeled data using contrastive tasks can significantly improve the model's performance on downstream speech tasks [6; 7] by providing useful compressed latent representation which is easier to model than raw waveforms or spectrograms.

These related methodologies highlight the breadth of techniques employed in modern TTS systems, from end-to-end models to sophisticated generative networks using embeddings and contrastive learning. The convergence of these technologies represents a significant step forward in the quest for more natural and expressive synthetic speech.

Diffusion Models. Diffusion models have recently been explored as a powerful method for generating high-quality speech. These models, such as WaveGrad and DiffWave, use a gradual denoising process to synthesize speech, starting from noise and progressively refining the signal into intelligible speech [8; 9]. The process involves a learned reverse diffusion that transforms a Gaussian noise distribution into a complex signal [10]. Transformer-based Text-to-Speech Models. Transformer-based architectures have significantly influenced the development of TTS systems, offering substantial improvements over traditional methods. These models fall into two main categories: autoregressive and non-autoregressive models, each with unique attributes and applications in speech synthesis. Autoregressive Models: Autoregressive models, such as [11], generate signals sequentially, predicting one segment at a time based on all previously generated segments. This approach ensures high coherence and naturalness in the speech output. The transformer's attention mechanism allows these models to capture long-range dependencies in text, crucial for prosody and intonation in speech. A typical example includes the original Transformer TTS, which utilizes a self-attention mechanism to model temporal sequences in a highly parallelizable manner:

$$p(y|x;\theta_{AR}) = \prod_{t=0}^{T} p(y_t|y_{<t}, x;\theta_{AR}) \tag{2.1}$$

where y_t is the predicted audio output at time t, and x_t is the input phoneme or text sequence up to time t.

Non-autoregressive Models: In contrast, non-autoregressive models such as FastSpeech [12] bypass the sequential dependency of autoregressive models, predicting all parts of the speech output simultaneously. This leads to significantly faster synthesis times and reduces latency, which is beneficial for real-time applications. FastSpeech and its successors, like FastSpeech 2 [13], improve on this approach by predicting duration, pitch, and energy explicitly, which are then used to modulate the speech synthesis process.

$$\hat{y} = \text{Parallel Decoder}\big(\text{Duration Predictor}(x)\big) \tag{2.2}$$

where $\hat{y}$ represents the entire speech waveform generated in parallel, and x is the input phonetic/text representation.

Both types of transformer-based TTS models have pushed the boundaries of speech synthesis, offering more natural, flexible, and efficient solutions. However, the choice between autoregressive and non-

autoregressive approaches often depends on the specific requirements of latency, naturalness, output diversity and computational resources.

3. Model description

Masked Autoencoder for Feature Compression. The Masked Autoencoder (MAE) processes an input audio signal x by first computing its log mel spectrogram $X \in R^{T\times F}$, where T indicates the time steps, and F represents the mel frequency bins. This spectrogram X is analogized to an image and segmented into patches of size $P \times P$, where each patch size P is a divisor of both T and F. These patches are then input into the AudioMAE encoder. The encoder, a convolutional neural network, operates with a kernel and stride both set to P, producing an output with D channels. Consequently, the encoder output is $E \in R^{T'\times F'\times D}$, where $T' = \frac{T}{P}$ and, $F' = \frac{F}{P}$ and D is the dimension of the embedding produced by MAE. The encoded features E are treated as the latent representation for subsequent processing.

To train the AudioMAE, a loss function is employed, specifically the Mean Squared Error (MSE) loss, calculated over the masked patches to assess the reconstruction quality. The MSE loss is defined as:

$$\text{MSE Loss} = \frac{1}{N_{\text{masked}}}\sum_{i=1}^{N_{\text{masked}}}\left(\widehat{X_i} - X_i\right)^2 \tag{3.1}$$

where N_{masked} is the number of masked patches, X_i is the original patch, and $\widehat{X_i}$ is the reconstructed patch output by the decoder of the MAE.

Conditioning Information C: Reference Audio and Text Phonemes. In the audio generation model, the conditioning information C plays a crucial role in guiding the generative process by providing contextual cues that influence the output. For this model, C is derived from two primary sources: reference audio and text phonemes, each contributing unique aspects to the generation process.

CLAP Autoencoder for Conditioning. The CLAP autoencoder is designed to project both audio and text into a unified multimodal space, enabling the effective use of this information as conditioning data. Let Xa denote the processed audio, represented in a matrix $X_a \in R^{F\times T}$, where F is the number of spectral components, such as Mel bins, and T is the number of time bins. Similarly, let X_t denote the text representation. Within a batch of N audio-text pairs, these are denoted as $\{X_a, X_t\}$.

The audio and text data are encoded via separate encoder functions, $f_a(\cdot)$ and $f_t(\cdot)$ respectively. For a batch of N items, the encoded representations are given by:

$$\widehat{X_a} = f_a(X_a); \quad \widehat{X_t} = f_t(X_t) \tag{3.2}$$

where $\widehat{X_a} \in R^{N\times V}$ and $\widehat{X_t} \in R^{N\times U}$ represent the dimensionalities V and U of the audio and text representations, respectively.

To bring these representations into a joint multimodal space of dimension d, learnable linear projections are applied:

$$E_a = L_a\left(\widehat{X_a}\right) \tag{3.3}$$

$$E_t = L_t\left(\widehat{X_t}\right) \tag{3.4}$$

where $E_a, E_t \in R^{N\times d}$ are the projected embeddings for audio and text, and L_a, L_t are the respective linear projection functions.

The similarity between the audio and text embeddings is computed in the joint space as follows:

$$C = \tau \cdot (E_t \cdot E_a^{\top}) \tag{3.5}$$

where τ is a temperature parameter that scales the range of the logits. The similarity matrix $C \in R^{N\times N}$ includes correct pairs along the diagonal and incorrect pairs off the diagonal.

A symmetric cross-entropy loss is then computed over the similarity matrix to train the encoders and their projections:

$$\mathcal{L} = 0.5 \cdot \left(\mathrm{l}_{\text{text}}(C) + \mathrm{l}_{\text{audio}}(C)\right) \tag{3.6}$$

where $\mathrm{l} = \frac{1}{N}\sum_{i=0}^{N}\log\left(\text{softmax}(C)\right)$ along the text and audio axes, respectively. This loss function facilitates the joint training of the audio and text encoders, enhancing their capability to encode relevant features effectively for audio generation tasks.

Text phoneme encoding: Text phonemes represent another vital component of the conditioning information. Phonemes, the smallest units of sound in a language, are extracted from the input text and encoded to capture the linguistic nuances and articulatory features necessary for generating coherent and contextually appropriate audio. This encoding process transforms textual data into a sequence of phonetic representations, C_{phonemes}, which are then used to condition the audio generation, ensuring that the produced audio matches the intended linguistic content and style dictated by the input text.

Together, these conditioning components $C = \{C_{\text{ref}}, C_{\text{phonemes}}\}$ integrate multiple modalities—audio and text—providing a comprehensive set of cues that enhance the model's ability to generate high-fidelity and contextually rich audio outputs.

Autoregressive Modeling for Intermediate Representation. This model component is responsible for generating a latent representation from diverse conditioning information using an autoregressive approach inspired by transformer-based models. The formulation of the autoregressive model $\mathcal{M}_\theta$ is given by:

$$\hat{Y} = \mathcal{M}_\theta(C) \tag{3.7}$$

where C represents conditioning information, and $\hat{Y}$ is the predicted latent representation. The model $\mathcal{M}_\theta$, parameterized by θ, predicts the next sequence element based on previous ones, maximizing the probability distribution across the sequence:

$$\text{argmax}_\theta \prod_{i=1}^{L} P\left(y_i \middle| C_{\text{ref}}, C_{\text{phonemes}}, y_1, y_2, \dots, y_{i-1}; \theta\right) \tag{3.8}$$

where L is the length of the latent sequence Y encoded by MAE, and y_i are its components. Variational Autoencoder (VAE) for Diffusion Modeling: A Variational Autoencoder (VAE) [14] primarily for feature compression and to learn a compact audio representation, z, which is dimensionally much smaller than the original audio signal, x.

The operation of the VAE can be expressed through the forward pass equation:

$$\mathcal{V}: X \mapsto z \mapsto \hat{X} \tag{3.9}$$

where X represents the mel-spectrogram of the audio input x, and $\hat{X}$ is the reconstruction of X. This reconstructed spectrogram, $\hat{X}$, can subsequently be transformed back into the audio waveform $\hat{x}$ using a pretrained HiFiGAN vocoder [15].

To optimize the parameters of the VAE, a reconstruction loss and a discriminative loss are computed based on the comparison between X and $\hat{X}$. Furthermore, the VAE architecture employs a regularization strategy by computing the KullbackLeibler (KL) divergence between the latent representation z and a standard Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$:

$$\text{KL Loss} = D_{\text{KL}}\left(\mathcal{N}(z; \mu_z, \sigma_z^2) \parallel \mathcal{N}(0,1)\right) \tag{3.10}$$

This regularization helps to maintain the statistical properties of the latent space, ensuring that z adheres closely to a Gaussian distribution, thereby stabilizing the generation process and enhancing the quality of the reconstructed audio.

Latent diffusion model for audio synthesis. The audio synthesis is performed using a latent diffusion model that operates within the latent space provided by the VAE autoencoder. This model is expressed through a series of diffusion steps, starting with a latent representation z and gradually adding noise to reach a diffusion state z_T:

$$z_t = \sqrt{1 - \beta_t} z_{t-1} + \sqrt{\beta_t} \epsilon_t \tag{3.11}$$

where β_t is a noise schedule parameter, and $\epsilon_t \sim \mathcal{N}(0, I)$ is Gaussian noise.

The reverse process involves a gradual denoising of z_T to reconstruct the latent representation:

$$z_{t-1} = \frac{z_t - \sqrt{\beta_t}\epsilon_t}{\sqrt{1-\beta_t}} \tag{3.12}$$

The optimization targets the minimization of the difference between the original and reconstructed latent representations, defined by the loss function:

$$\mathcal{L}(\phi) = E_{z_{\mathbb{0}}, \epsilon \sim \mathbb{N}(\mathbb{0}, I), t}[|z_0 - \text{Dec}(z_t; \phi)|^2] \tag{3.13}$$

where ϕ are the parameters of the diffusion model, Dec denotes the decoding function of the diffusion model, and z_0 is the original latent representation.

Adaptation of Pretrained Model Components. In the development of our model, we utilized components from the pretrained Audio Latent Diffusion Model 2 (Audio LDM2). This approach allowed us to leverage the robust foundations established by the existing model, particularly its effective handling of complex audio data through diffusion processes. An important modification in our methodology involved the adaptation of the conditioning mechanism used in Audio LDM2. Traditionally, Audio LDM2 employs a conditioning vector C that incorporates CLAP-encoded text embeddings to guide the audio synthesis process. In contrast, our model replaces these text embeddings with CLAP-encoded audio embeddings which is intended to encode emotion and speaker information. This change aligns better with our focus on enhancing audio quality and relevance in text-to-speech applications, where the direct correlation between the input audio characteristics and the generated output is crucial.

$$C_{\text{ref}} = f_a(X_{ref}) \tag{3.14}$$

where C_{ref} represents the new conditioning vector using audio embeddings, $f_a(\cdot)$ is the CLAP audio encoder, and X_{ref} is the referebce audio feature matrix.

This adaptation not only tailors the model to our specific use case more closely, but also optimizes the interaction between the conditioning information and the generative components of the model. By integrating audio embeddings directly, our model gains a more nuanced understanding of the audio features, potentially leading to more accurate and lifelike audio generation in TTS systems.

4. Evaluation metrics

This section describes the evaluation process of our generative model.

The performance of the updated AudioLDM2 model is evaluated using several key metrics:

- Speaker Similarity: Quantifies the ability of the TTS system to preserve the unique characteristics of the speaker's voice.
- Emotion Classification Error: Measures the model's accuracy in conveying the intended emotional states in the synthesized speech.
- Word Error Rate (WER) / Character Error Rate (CER): Assesses the intelligibility and accuracy of the spoken output, comparing the transcribed text from the synthesized speech to the original input text.

All metrics reported below were calculated using the Amphion software [16] - a toolkit library for audio generation. These metrics provide a comprehensive framework for assessing the effectiveness of the TTS system in producing high-quality, emotionally expressive, and speaker-specific speech.

5. Results and metrics description

Speaker Similarity Metric. To quantitatively assess the speaker similarity between the reference and generated audio samples, we employed a speaker verification model based on WavLM, a state-of-the-art audio processing model [7]. This model was pretrained using a contrastive loss, which optimizes the embeddings to minimize the distance between similar pairs and maximize the distance for dissimilar pairs, making it well-suited for speaker verification tasks.

The metric we report is the average cosine similarity between the embeddings of reference audio samples and their corresponding generated samples. The embeddings are extracted using the WavLM model, which captures speaker-specific characteristics. Cosine similarity measures the cosine of the angle between two vectors in the embedding space, providing a scale from -1 (completely different) to 1 (identical), where higher values indicate greater speaker similarity. The formula for cosine similarity is given by:

$$\text{Cosine Similarity} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.1}$$

where Ai and Bi are the components of the embeddings from the reference and generated audio samples, respectively. This metric effectively quantifies how well the generated audio preserves the identity characteristics of the speaker in the reference audio. The results are presented in Table 1

Table 1. Comparison of speaker similarity scores

Model	Speaker similarity
Proposed model	0.63
xTTS v2	0.9

Emotional Classification Accuracy. The accuracy of emotion recognition was evaluated using the Emotion2Vec model [17], which predicted emotions for both the reference and the produced audios. This measure reflects the model's ability to encode and reproduce the emotional states intended by the original speech. Results are tabulated in Table 2.

Table 2. Comparison of emotional classification accuracy

Model	Emotion classification accuracy
Proposed model	0.035
xTTS v2	0.17

Metrics for measuring WER and CER were calculated using transcripts generated by a pre-trained large Whisper [18] Automatic Speech Recognition (ASR) model, comparing these against the ground truth transcripts.

WER is computed as the ratio of the total number of operations (insertions, deletions, and substitutions) needed to convert the ASR-generated transcript into the ground truth transcript, divided by the total number of words in the ground truth transcript. The formula for WER is given by:

$$\text{WER} = \frac{S+D+I}{N} \tag{5.2}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the ground truth transcript.

Similarly, CER is calculated by applying the same principle at the character level rather than the word level. It measures the minimum number of insertions, deletions, and substitutions required to change the ASR-generated transcript into the ground truth, normalized by the total number of characters in the ground truth transcript. The formula for CER is:

$$\text{CER} = \frac{s+d+i}{n} \tag{5.3}$$

where s represents substitutions, d represents deletions, i represents insertions, and n is the total number of characters in the ground truth transcript.

Both metrics provide crucial insights into the transcription accuracy of the generated speech, with lower values indicating higher accuracy and better performance of the text-to-speech synthesis system. The results are presented in Table 3.

Table 3. Comparison of Word error rate and Character error rate

Model	Word error rate↓	Character error rate↓
Proposed model	1.0	1.01
xTTS v2	0.21	0.02

5. Conclusions

This study provided the evaluation of the latent diffusion model, against the xTTS v2 model using a set of rigorous metrics on the EmoV-DB dataset. The findings revealed some insights into the performance of both models in terms of speaker similarity, emotional preservation, and intelligibility.

While the proposed model demonstrated decent efficiency in maintaining speaker characteristics, as indicated by the speaker similarity score, it was outperformed by xTTS v2 in all assessed metrics. Notably, our model exhibited considerable deficiencies in emotional classification accuracy, suggesting that the audio CLAP embeddings it relies on may be more attuned to capturing speaker-related information than the nuances of emotional expression. This observation was underscored by the model's

low emotion classification error rate, which points to a potential misalignment between the emotional intents encoded by the embeddings and those expressed in the speech output. Also, pretrained dataset for the CLAP component, which is a mix of speech and general audio and its corresponding captions, might not be effective for speech synthesis, suggesting that pre-training on transcribed speech dataset may improve generation quality.

REFERENCES

1. Y. Wang et al. Tacotron: Towards End-to-End Speech Synthesis. *Interspeech 2017. ISCA: ISCA*, 20-24 August 2017, Stockholm, Sweeden, 2017, p. 4006-4010
2. J. Shen et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 15–20 April 2018, Calgary, AB, Canada, 2018, p. 4779-4783
3. Taigman Y. Voiceloop: Voice fitting and synthesis via a phonological loop, 2018 (Preprint Arxiv:1707.06588)
4. Lee Y. Emotional end-to-end neural speech synthesizer, 2017 (Preprint Arxiv: 1711.05447)
5. Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick and S. Dubnov. Large-Scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023 - 2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4–10 June 2023, Rhodes Island, Greece. 2023
6. W.-N. Hsu et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*. 2021. vol. 29. p. 3451–3460.
7. S. Chen et al. WavLM: large-scale self-supervised pre-training for full stack speech processing. *IEEE journal of selected topics in signal processing*. 2022. Vol. 16, p. 1505–1518.
8. Chen N. Wavegrad: Estimating gradients for waveform generation. *International Conference on Learning Representations (ICLR)*, 2020.
9. Kong Z. Diffwave: A versatile diffusion model for audio synthesis. *International Conference on Learning Representations (ICLR)*, 2021.
10. Chen M. An overview of diffusion models: Applications, guided generation, statistical rates and optimization, 2024 (Preprint Arxiv: 2404.07771)
11. Wang C. Neural codec language models are zero-shot text to speech synthesizers, 2023 (Preprint Arxiv: 2301.02111)
12. Ren Y. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 8-14 December 2019, Vancouver Convention Centre, Canada, vol 32.
13. Ren Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. *ICLR 2021 The Ninth International Conference on Learning Representations*, 2021
14. Kingma D. P. Auto-encoding variational bayes. *International Conference on Learning Representations*. 14-16 April 2014, Banff, AB, Canada, 2014
15. Kong J. et al. Hifi-gan: Generative adversarial networks for efficient and high-fidelity speech synthesis. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 6-12 December, 2020, vol 33, p. 17022-17033.
16. Xueyao Zhang, Liumeng Xue, Yicheng Gu ett.al. Amphion: An open-source audio, music and speech generation toolkit, 2024 (Preprint Arxiv: 2312.09911)
17. Ma Z. et al. Emotion2vec: Self-supervised pre-training for speech emotion representation, 2023 (Preprint Arxiv: 2312.15185)
18. Radford A. Robust speech recognition via large-scale weak supervision. *Proceedings of Machine Learning Research*, 23-29 July, 2023, vol 202, p. 28492-28518.

Іванюк Андрій Олегович *Аспірант докторської школи*
Національний університет "Києво-Могилянська академія", факультет інформатики, вулиця Григорія Сковороди 2, Київ, Україна, 04655
e-mail: a.ivaniuk@ukma.edu.ua
https://orcid.org/0000-0002-4189-3787

Модель латентної дифузії для обробки мовного сигналу

Актуальність. Розробка генеративних моделей для синтезу аудіо, включаючи текст-у-мовлення (англ. text-to-speech, TTS), текст-у-музику та текст-у-аудіо застосування, значною мірою залежить від їх здатності обробляти складні та різноманітні вхідні дані. В цій роботі ми розглядаємо латентне дифузійне моделювання - універсальний підхід, який використовує стохастичні процеси для генерації високоякісних аудіо сигналів.

Мета. Це дослідження має на меті оцінити ефективність латентного дифузійного моделювання для аудіо синтезу на основі набору даних EmoV-DB, який містить записи з багатьма мовцями, з п'ятьма емоційними станами, та порівняти його з іншим генеративними методом.

Методи дослідження. Ми застосували латентне дифузійне моделювання спеціально для синтезу мовлення та оцінили його ефективність за допомогою метрик, які визначають зрозумілість, подібність голосу та збереження емоцій в згенерованому аудіо сигналі.

Результати. Дослідження показує, що запропонована модель демонструє пристойну ефективність у збереженні характеристик голосу, але поступається дискретній авторегресивній моделі: xTTS v2 за всіма оціненими метриками. Зокрема, досліджувана модель виявляє недоліки в точності класифікації емоцій, що вказує на можливе невідповідність між емоційними намірами, закодованими у векторах, та тими, що виражені у згенерованому сигналі.

Висновки. Результати вказують на те, що подальше вдосконалення здатності нейронної мережі кодувальника обробляти та інтегрувати емоційні дані покращує ефективність латентної дифузійної моделі. В наших подальших дослідженнях ми плануємо зосередитися на оптимізації балансу між характеристиками мовця та емоційними характеристиками в TTS моделях для досягнення більш цілісного та ефективного синтезу людського мовлення.

Ключові слова: аудіо моделювання, штучні нейронні мережі, синтез мовлення.

Надійшла у першій редакції 02.03.2024, в останній - 11.04.2024.
The first version has been received on 02.03.2024, the final version - on 11.04.2024.