СТВОРЕННЯ СИНТЕТИЧНИХ ГОЛОСІВ ДЛЯ АУДІОКНИГ

***Супрун О.О.**, к.т.н., доцент, кафедра МСТ, ХНУРЕ*
***Білова А.С.**, бакалавр, кафедра МСТ, ХНУРЕ*

***Анотація.** Це дослідження висвітлює створення синтетичних голосів для аудіокниг з використанням архітектур TTS на основі трансформерів та дифузійних моделей. У роботі проаналізовано перехід до автоматизованих процесів, акцентуючи увагу на підвищенні швидкості виробництва, мультимовній локалізації та доступності. Крім того, у статті розглянуто технічні обмеження просодії та етичні ризики, пов'язані з клонуванням голосу та інформаційною безпекою.*

***Ключові слова:** синтез мовлення, текст-у-мовлення (TTS), аудіокниги, нейронні мережі, клонування голосу, штучний інтелект.*

Розвиток штучного інтелекту суттєво змінив підходи до створення аудіоконтенту. Одним із найбільш динамічних напрямів є синтез мовлення, який дозволяє автоматично перетворювати текст у природне аудіо. У контексті аудіокниг це відкриває можливості швидкого та масштабованого виробництва контенту без необхідності залучення професійних дикторів. У сучасних рішеннях домінують нейронні архітектури, зокрема Transformer-based TTS моделі, які використовують механізм self-attention для обробки довгих послідовностей тексту. Такі моделі забезпечують високу контекстуальну узгодженість мовлення, мінімізуючи артефакти та синтаксичні помилки. Альтернативним підходом є використання RNN (Recurrent Neural Networks) та LSTM-архітектур, однак вони поступово витісняються більш ефективними трансформерами.

Важливим компонентом сучасних систем є нейронні вокодери, такі як WaveNet, WaveGlow, HiFi-GAN та Diffusion-based vocoders. Вони відповідають за генерацію часової форми аудіосигналу (waveform synthesis) з проміжних акустичних представлень, наприклад, mel-spectrograms. Саме вокодер визначає фінальну якість звучання, включаючи тембр, природність шумів та спектральну деталізацію.

Особливу роль у розвитку TTS відіграють дифузійні моделі (Diffusion Models), які дозволяють поступово відновлювати аудіосигнал через ітеративний процес денойзингу. Це забезпечує високу якість генерації мовлення з мінімальними спотвореннями та покращеною стабільністю інтонаційних характеристик.

Сучасні системи також інтегрують механізми speaker embedding та voice cloning, які дозволяють моделювати унікальні характеристики голосу конкретного диктора. Для цього використовуються embeddings у латентному просторі, отримані за допомогою speaker verification моделей (наприклад, d-vector, x-vector). Це дає змогу створювати кастомізовані голоси або повністю відтворювати голосову ідентичність на основі невеликої вибірки аудіоданих (few-shot learning).

У сфері аудіокниг синтетичне мовлення реалізується через pipeline автоматизованої генерації контенту, який включає: preprocessing тексту, segmentation на речення та абзаци, TTS inference, post-processing аудіо (normalization, loudness equalization, noise shaping). Часто застосовуються стандарти LUFS (Loudness Units relative to Full Scale) для вирівнювання гучності відповідно до вимог аудіоплатформ.

Практичне застосування технологій TTS у виробництві аудіокниг охоплює кілька ключових сценаріїв. Перш за все, це автоматизоване озвучення великих текстових корпусів із використанням batch processing. Другим напрямом є мультимовна локалізація (multilingual TTS), яка дозволяє швидко адаптувати контент для різних мовних аудиторій. Третім є accessibility-сегмент, орієнтований на користувачів із порушеннями зору, де синтетичне мовлення виступає як assistive technology.

Окремо слід відзначити інтеграцію TTS у real-time systems, де генерація мовлення відбувається у режимі низької затримки (low-latency inference). Це критично важливо для інтерактивних аудіокниг, освітніх платформ та VR/AR середовищ, де аудіо синхронізується з візуальними або інтерактивними елементами.

З точки зору інженерної реалізації, сучасні TTS-системи часто розгортаються у вигляді API-based cloud services із використанням GPU-обчислень для прискорення inference. Також застосовуються оптимізації, такі як model quantization, pruning та knowledge distillation, що дозволяють зменшити обчислювальну складність без суттєвої втрати якості.

Попри значний технологічний прогрес, синтез мовлення має низку суттєвих обмежень, які проявляються на різних рівнях архітектури та застосування (табл. 1). Одним із ключових технічних викликів є складність моделювання природної просодії мовлення, зокрема точного відтворення інтонаційних контурів, динаміки наголосу та варіативності темпу. Незважаючи на використання сучасних нейронних архітектур, таких як Transformer-based TTS та дифузійні моделі, повна емоційна репрезентація людського голосу залишається частково обмеженою.

Таблиця 1 – Порівняльний аналіз підходів до озвучення аудіокниг

Параметр	**Традиційне озвучення**	**Синтетичний голос (AI)**
Швидкість створення	Низька (дні/тижні)	Висока (хвилини/години)
Вартість	Висока	Низька
Масштабування	Обмежене	Практично необмежене
Емоційність мовлення	Висока	Середня / висока (залежить від моделі)
Гнучкість редагування	Низька	Висока
Мовна підтримка	Обмежена	Широка (багатомовність)
Персоналізація голосу	Обмежена	Висока

Додатковим технічним фактором є залежність якості синтезу від обсягу та репрезентативності навчальних даних. Нерівномірність корпусів мовлення, відсутність балансування за емоційними станами, акцентами та мовними стилями призводять до деградації якості генерації, зокрема появи артефактів мовлення, роботизованого тембру або нестабільної інтонації.

Етичні аспекти синтезу мовлення є не менш критичними. Однією з основних проблем є ризик несанкціонованого використання голосової ідентичності, що може призводити до створення deepfake-аудіо. Це формує нові виклики у сфері інформаційної безпеки, зокрема у контексті аудіодезінформації та маніпулятивного контенту. У зв’язку з цим актуальним стає впровадження механізмів цифрового маркування синтетичного мовлення, які дозволяють ідентифікувати походження згенерованого аудіо.

Соціальні наслідки впровадження технологій синтезу мовлення проявляються у трансформації ринку праці в креативних індустріях. Зокрема, спостерігається потенційне зменшення попиту на традиційних дикторів та озвучувачів, що змінює структуру зайнятості у сфері аудіопродакшену. Водночас виникають нові професійні ролі, такі як voice model trainer, AI audio engineer та synthetic media designer, що свідчить про перерозподіл компетенцій у цифровій економіці. Таким чином, можна стверджувати, що синтетичні голоси є невід'ємною складовою процесу цифрової трансформації медіаіндустрії. Вони забезпечують суттєве підвищення ефективності виробництва аудіоконтенту, розширюють можливості масштабування та сприяють підвищенню доступності інформації для різних категорій користувачів, включаючи осіб з інвалідністю.

Важливою перевагою є також підвищення рівня доступності інформації, зокрема для користувачів із порушеннями зору або іншими обмеженнями сприйняття текстового контенту. У цьому контексті синтетичне мовлення виступає як інструмент інклюзивних технологій (assistive technologies), що сприяє розширенню освітніх і комунікаційних можливостей для різних соціальних груп. Разом із тим інтеграція систем синтезу мовлення вимагає системного та відповідального підходу до їх використання. Це передбачає врахування як інженерних обмежень моделей (зокрема, якості генерації, стабільності просодичних характеристик та обчислювальної складності), так і етичних викликів, пов'язаних із використанням голосових даних, авторськими правами та потенційними ризиками зловживання технологіями voice cloning і deepfake-аудіо.

У перспективі подальший розвиток нейронних архітектур синтезу мовлення буде пов'язаний із вдосконаленням мульти-модальних генеративних моделей, які інтегрують текстові, візуальні та аудіальні модальності в єдину систему контекстного аналізу. Особливе значення матимуть контекстно-адаптивні TTS-системи, здатні динамічно змінювати параметри мовлення залежно від семантики тексту, комунікативної ситуації та емоційного контексту.

Розвиток таких технологій поступово наближає синтетичне мовлення до рівня природної людської комунікації, де ключовими характеристиками стають не лише акустична якість сигналу, а й когнітивна узгодженість, емоційна виразність та контекстна адаптивність.

Література.

1. Maddison, C.J., Tarlow, D., & Minka, T. (2014). A* sampling. Advances in Neural Information Processing Systems, 3086-3094.

2. Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5206-5210.

3. Pratap, V., Hannun, A., Xu, Q., Cai, J., Kahn, J., Synnaeve, G., Liptchinsky, V., & Collobert, R. (2019). Wav2letter++: A fast open-source speech recognition system. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).