

ПОРІВНЯЛЬНИЙ АНАЛІЗ СИСТЕМ ДЛЯ АВТОМАТИЗОВАНОГО ОЗВУЧЕННЯ ВІДЕО

Донецький національний університет імені Василя Стуса, м. Вінниця

Системи синтезу мовлення (TTS) відіграють ключову роль в автоматизованому озвученні текстових сценаріїв, особливо у відеопродукції, де швидкість синтезу та природність голосу є надзвичайно важливими. Сучасні нейронні моделі, як-от Tacotron 2 та FastSpeech 2, у поєднанні з вокодерами на кшталт WaveNet демонструють значний прогрес у генерації мовлення, близького до людського [1].

Метою цієї роботи є проведення всебічного порівняльного аналізу сучасних систем синтезу мовлення (TTS), які знаходять застосування в автоматизованому озвученні відеоконтенту. Особлива увага приділяється дослідженню їх архітектурних особливостей, фундаментальних принципів функціонування, а також оцінці якості та швидкості генерації мовлення. До того ж важливим аспектом є виявлення ключових тенденцій, що визначають подальший розвиток технологій у цій динамічній галузі.

Для досягнення поставленої мети необхідно вирішити комплекс завдань. По-перше, потрібно детально дослідити ключові архітектури провідних TTS-систем, як-от Tacotron 2 та FastSpeech 2, включно з аналізом їх основних компонентів, зокрема мереж прогнозування акустичних ознак та вокодера. По-друге, важливо проаналізувати роль та принципи роботи вокодерів, зокрема WaveNet та універсальних нейронних вокодерів, у контексті забезпечення високої природності синтезованого мовлення. По-третє, необхідно розглянути сучасні підходи, спрямовані на прискорення процесу синтезу мовлення, що включають вивчення неавто-регресійних моделей на прикладі FastSpeech 2 та повністю наскрізних систем, як-от FastSpeech 2s. Наступним кроком є проведення порівняльної характеристики розглянутих TTS-систем за низкою ключових критеріїв: тип вхідних даних, наявність проміжного представлення, принцип роботи вокодера, тип генерації (авто-регресійний чи неавто-регресійний) та їх основні переваги. На завершення, на основі проаналізованих технологій, потрібно визначити основні вектори розвитку систем автоматизованого озвучення відео.

Більшість передових TTS-систем, наприклад, Tacotron 2, використовують двокомпонентну архітектуру: мережу прогнозування акустичних ознак (часто мел-спектрограм) з тексту та вокодер, який синтезує з цих ознак звукову хвилю. Мережа прогнозування зазвичай базується на архітектурі sequence-to-sequence з механізмом уваги, що забезпечує точне вирівнювання тексту та мовлення. Мел-спектрограми, як проміжне представлення, спрощують тренування та є інваріантними до фази [2].

Для досягнення високої природності звучання вирішальну роль відіграють вокодери. WaveNet, авторегресивна генеративна модель, що працює безпосеред-

ньо з сирими аудіоданими, використовує дилатовані каузальні згортки для моделювання довготривалих залежностей, забезпечуючи високу якість синтезу [2–3]. Існують також універсальні нейронні вокодери на базі WaveRNN, здатні працювати з багатьма дикторами та мовами [4].

Для подолання повільності авторегресійних моделей розробляються системи для швидкого синтезу, як-от FastSpeech 2. Вони відмовляються від авторегресії на етапі генерації мел-спектрограм, тренуються на реальних даних та інтегрують додаткову інформацію про варіативність мовлення (висота тону, енергія, тривалість фонем) для кращого вирішення проблеми «один-до-багатьох» [5].

Тенденцією в розвитку TTS-систем є перехід до повністю наскрізних моделей, як FastSpeech 2s, що генерують звукову хвилю безпосередньо з тексту, мінаючи проміжне представлення. Це спрощує архітектуру та потенційно прискорює синтез, часто із застосуванням адверсаріального навчання для покращення якості [5]. У табл. 1 наведено порівняльну характеристику TTS-систем.

Таблиця 1 – Порівняльна характеристика ключових TTS-систем

Характеристика	Tacotron 2	FastSpeech 2s	WaveNet (як вокодер)
Вхідні дані	Ембединги символів	Послідовність фонем	Мел-спектрограми або лінгвістичні ознаки
Проміжне представлення	Мел-спектрограми	Прямий синтез хвилі	Відсутнє (генерує звукову хвилю)
Принцип вокодера	Модифікований WaveNet	Внутрішній генератор хвилі	Авторегресійна генерація сирого аудіо
Тип генерації (спектрограма)	Авторегресійний (Seq2Seq + увага)	Неавторегресійний	Не застосовується
Тип генерації (хвиля)	Авторегресійний (WaveNet)	Неавторегресійний	Авторегресійний
Ключові переваги	Висока природність, MOS близький до людського мовлення	Дуже швидкий синтез, висока якість, вирішення проблеми «один-до-багатьох» через використання адаптерів варіативності (тон, енергія, тривалість)	Найвища природність генерації звукової хвилі, слугує еталоном якості

Отже, проведений аналіз систем для автоматизованого озвучення відео демонструє значний прогрес у галузі синтезу мовлення, зумовлений переважно розвитком нейронних мереж. Системи, як-от Tacotron 2, що базуються на двокомпонентній архітектурі з генерацією мел-спектрограм та використанням вокодерів на кшталт WaveNet, досягають високої природності звучання, наближеної до людського мовлення. Водночас для вирішення проблеми швидкості синтезу, критично важливої для відеопродукції, були розроблені неавторегресійні моделі, як FastSpeech 2, що забезпечують швидку генерацію за умови збереження високої якості та ефективного моделювання варіативності мовлення. Тенденція до створення повністю наскрізних систем, як FastSpeech 2s, що генерують звукову хвилю безпосередньо з тексту, свідчить про прагнення до спрощення архітектур та подальшого прискорення процесу синтезу.

З огляду на дані можна зробити висновок, що якщо пріоритетом є максимальна природність голосу, а час синтезу не є критичним обмеженням, Tacotron 2 у

поєднанні з високоякісним вокодером (як WaveNet) може вважатися кращим вибором, оскільки він часто встановлює стандарт якості. Проте якщо ключовими факторами є швидкість генерації, ефективність і можливість обробки великих обсягів тексту для відео, то системи на кшталт FastSpeech 2 та особливо FastSpeech 2s є значно привабливішими. Вони пропонують відмінний компроміс між швидкістю та якістю, що робить їх більш практичними для багатьох реальних завдань автоматизованого озвучення. Отже, оптимальна TTS-система обирається шляхом зважування переваг у природності звучання проти переваг у швидкості та ефективності генерації відповідно до потреб конкретного проекту.

Список використаних джерел

1. A Survey on Neural Speech Synthesis / X. Tan, T. Qin, F. Soong, T. -Y. Liu. *arXiv.org*. 2021. URL: <https://arxiv.org/abs/2106.15561> (дата звернення: 12.05.2025).
2. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions / J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu. *arXiv.org*. 2017. URL: <https://arxiv.org/abs/1712.05884> (дата звернення: 12.05.2025).
3. WaveNet: A Generative Model for Raw Audio / A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu. *arXiv.org*. 2016. URL: <https://arxiv.org/abs/1609.03499> (дата звернення: 12.05.2025).
4. Towards achieving robust universal neural vocoding / J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, V. Aggarwal. *arXiv.org*. 2018. URL: <https://arxiv.org/abs/1811.06292> (дата звернення: 12.05.2025).
5. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech / Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu. *arXiv.org*. 2020. URL: <https://arxiv.org/abs/2006.04558> (дата звернення: 12.05.2025).

УДК 004.415.2:004.657

*Лисак В. О., здобувач вищої освіти 2 курсу
спеціальності 122 Комп'ютерні науки
Зелінська О. В., канд. техн. наук, доцент,
в. о. завідувача кафедри інформаційних технологій*

ОПТИМІЗАЦІЯ ЗАПИТІВ ДО БАЗИ ДАНИХ У ВЕБПРОЄКТАХ

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасних вебпроєктах, де PHP виступає серверною мовою, а MySQL – системою керування базами даних, обсяг інформації постійно зростає. Відповідно ефективність SQL-запитів стає критично важливою для продуктивності системи. Запити без належної оптимізації можуть значно «стиснути» швидкодію програми та спричинити затримки у відповіді [1].

Для забезпечення взаємодії між вебдодатками та базами даних використовуються переважно два підходи: безпосереднє виконання SQL-запитів або застосування ORM (Object-Relational Mapping). Перший підхід полягає у прямому написанні SQL-команд у коді програми для роботи з базою даних. Натомість другий підхід ORM представляє технологію абстракції, що забезпечує відображення структур бази даних на об'єкти цільової мови програмування вебдодатку [7].