УДК 004.93 DOI: 10.31673/2786-8362.2025.013553

Іщеряков С.М., к.т.н.; **Попов А.О.**

ГЛИБОКЕ НАВЧАННЯ В СФЕРІ СИНТЕЗУ МОВЛЕННЯ

Ishcheriakov S.M., Popov A.O. Deep learning in speech synthesis systems. Deep learning systems allow you to automate complex tasks that previously required human intelligence, and do so with high accuracy. Deep learning uses artificial neural networks with many layers – each layer processes information in an increasingly complex and abstract way. This allows the system to learn high-level features such as emotions, intonations, expressiveness, etc. The introduction of these features makes synthesized speech more natural, which contributes to its better perception by humans. Unlike traditional speech synthesis methods, such as formant synthesis, concatenative synthesis or HMM-based approaches (Hidden Markov Models), deep learning provides much higher flexibility and sound quality. In traditional systems, speech was built from pre-recorded fragments or generated according to predefined rules, which limited the naturalness, intonation richness and emotional coloring of the voice. Thus, deep learning overcomes key limitations of traditional approaches and opens up new opportunities in the field of voice technologies – from text-to-speech to full-fledged emotional communication between humans and machines. The article considers the main areas of application of deep learning for speech synthesis, analyzes existing approaches to building synthesis systems, and analyzes their weaknesses and strengths.

Keywords: deep learning, neural network, synthesized speech

Іщеряков С.М., Попов А.О. Глибоке навчання в сфері синтезу мовлення. Системи глибокого навчання дозволяють автоматизувати складні завдання, які раніше вимагали людського інтелекту, і роблять це з високою точністю. Глибоке навчання використовує штучні нейронні мережі з багатьма шарами – кожен шар обробляє інформацію дедалі складніше й абстрактніше. Це дозволяє системі навчатися високорівневими ознакам, таким як емоції, інтонації, виразність. тощо. Впровадження цих ознак робить синтезоване мовлення більш природним, що сприяє кращому сприйняттю його людиною. На відміну від традиційних методів синтезу мовлення, таких як формантний синтез, конкатенативний синтез або HMM-основні підходи (Hidden Markov Models), глибоке навчання забезпечує значно вищу гнучкість і якість звучання. У традиційних системах мовлення будувалося з наперед записаних фрагментів або генерувалося за заздалегідь визначеними правилами, що обмежувало природність, інтонаційне багатство та емоційне забарвлення голосу. Таким чином, глибоке навчання долає ключові обмеження традиційних підходів і відкриває нові можливості в галузі голосових технологій – від тексту в мовлення до повноцінної емоційної комунікації між людиною і машиною. У статті розглянуто основні напрямки застосування глибокого навчання для синтезу мовлення, проведено аналіз існуючих підходів в побудові систем синтезу та проаналізовано їх слабкі та сильні сторони.

Ключові слова: глибоке навчання, нейронна мережа, синтез мовлення

Вступ

Глибоке навчання (англ. Deep Learning) – це підрозділ машинного навчання, який базується на штучних нейронних мережах із великою кількістю шарів. Його основна ідея – моделювання роботи людського мозку для розпізнавання складних шаблонів в даних [8]. Це спосіб навчити комп'ютер розуміти зображення, текст, мову чи звук без явного програмування, використовуючи великі обсяги даних і багатошарові нейронні мережі. Глибоке навчання продемонструвало видатні результати в різних сферах, зокрема в розпізнаванні мовлення, машинному перекладі, генерації тексту та синтезі голосу. Сучасні моделі, такі як трансформери та глибокі рекурентні нейронні мережі, здатні аналізувати складні залежності в мовленні та забезпечувати точність і природність взаємодії між людиною і машиною. Це відкриває нові можливості для створення інтелектуальних систем, які можуть розуміти контекст, емоції та наміри користувача, що є критично важливим для розвитку персональних асистентів, сервісів автоматичного перекладу та адаптивних освітніх платформ.

Вивчаючи ключові компоненти та методи, що використовуються в цих архітектурах, ми прагнемо надати уявлення про поточний стан глибокого навчання у сфері обробки мовлення та пролити світло на перспективи, які він відкриває для майбутніх досягнень у цій галузі.

Постановка завдання. У зв'язку з активним розвитком технологій глибокого навчання, зокрема в області обробки природної мови, виникає необхідність у систематичному аналізі сучасних архітектур і методів, що застосовуються для обробки мовлення. Зокрема, важливо дослідити переваги та недоліки різних підходів, таких як рекурентні нейронні мережі, двонаправлені RNN, вокодери, моделі на основі DNN, DBN, LSTM та seq2seq. Завданням дослідження є узагальнення наукових досягнень у галузі обробки мовлення з використанням глибокого навчання та окреслення перспектив подальших досліджень для підвищення точності, якості та ефективності синтезу й розпізнавання мовлення.

Аналіз останніх досліджень. За останні роки дослідження в галузі глибокого навчання для обробки мовлення досягли значних результатів завдяки розвитку новітніх архітектур, таких як трансформери, моделі з механізмами уваги (attention), та рекурентні нейронні мережі (RNN). Це дозволило значно покращити точність розпізнавання мовлення та синтезу голосу, особливо в умовах багатомовного середовища. Зокрема, модель Wav2Vec 2.0, розроблена компанією Facebook AI, продемонструвала високу ефективність у навчанні на сирих аудіо даних без необхідності використання великих обсягів розмічених даних, що є важливим кроком у напрямку самовільного навчання [1].

Інші підходи, такі як HuBERT, покладаються на кластеризацію та masked prediction, що дозволяє моделі адаптуватися до нових мовних контекстів без потреби в значних обсягах розмічених даних. Ці моделі досягли конкурентоспроможних результатів з точки зору точності порівняно з Wav2Vec 2.0, однак вони мають складнішу попередню обробку, що може стати перешкодою для їх широкого застосування [2].

Прогресивні системи, зокрема Whisper від OpenAI, поєднують можливості перекладу, транскрибування та виявлення мов, що значно підвищує універсальність моделей для багатомовних задач. Однак, зважаючи на їх високу точність, вони потребують значних обчислювальних ресурсів, що обмежує їх використання в реальних умовах [3].

Концепції, такі як Conformer, які поєднують в собі можливості згорткових мереж (CNN) та трансформерів, також набувають популярності, оскільки вони дозволяють досягти балансу між точністю виявлення локальних та глобальних патернів в мовленні. Однак, складність архітектури цих моделей потребує великої кількості обчислювальних потужностей і може бути обмеженням для їх реалізації в деяких сценаріях [4].

Системи, засновані на механізмах уваги, наприклад Tacotron або Char2Wav, з кожним роком показують все кращі результати у синтезі мовлення. Вони забезпечують високу якість без використання ручного маркування, що робить їх більш адаптивними для широкого кола додатків, таких як персональні асистенти та автоматичні перекладачі [5].

Проаналізувавши останні наукові публікації можна зробити висновок, що загалом, хоча досягнуто значного прогресу в обробці мовлення завдяки глибокому навчання, є й певні виклики, що залишаються актуальними: нестабільність моделей при навчанні, потреба в адаптації до індивідуальних голосів та складність обробки великих мовних конструкцій. Усі ці питання є об'єктом активних наукових досліджень, що передбачає подальші інновації в галузі глибокого навчання для обробки мовлення.

Метою роботи є аналіз сучасних архітектур глибокого навчання, що застосовуються в задачах обробки мовлення, з метою виявлення їхнього потенціалу, ефективності та впливу на якість мовленнєвих технологій. Особлива увага приділяється методам, які дозволяють моделювати складні контексти, відображати зв'язки між акустичними параметрами та підвищувати природність синтезованого мовлення.

Виклад основного матеріалу дослідження

Архітектури глибокого навчання в обробці мовлення: сучасний стан і перспективи розвитку. У цьому дослідженні розглядаються ключові компоненти та методи, що використовуються в архітектурах глибокого навчання, з метою окреслення сучасного стану цієї технології в контексті обробки мовлення та визначення перспектив її подальшого розвитку. Архітектури глибокого навчання здійснили суттєвий прорив у галузі автоматичної

обробки мови, продемонструвавши високу ефективність у виконанні низки завдань, таких як автоматичне розпізнавання мовлення, ідентифікація мовця та синтез мовлення. Завдяки здатності формувати ієрархічні уявлення з необроблених мовних даних, ці моделі суттєво перевершують традиційні підходи, що базуються на ручному виокремленні ознак. Архітектури глибокого навчання забезпечують ефективне виявлення складних патернів, розкриття латентних характеристик мовних сигналів та витягнення релевантної інформації з великих обсягів мовних даних. У межах даного дослідження проаналізовано основні типи нейронних архітектур, що застосовуються у завданнях обробки мовлення, окреслено їхні переваги, недоліки та вплив на розвиток галузі, таблиця 1.

Таблиця 1

Порівняння моделей генерації на основі глибокого навчання з класичними моделями

Тип	Приклади моделей	Переваги	Недоліки
Класичні моделі	Hidden Markov Models (HMM) Gaussian Mixture Models (GMM) n-грамні моделі (n-gram LM) Dynamic Time Warping (DTW)	Простота реалізації Невеликі обчислювальні витрати Добре працюють на вузькоспеціалізованих або контрольованих датасетах Добре підходять для реального часу при обмежених ресурсах	Слабка здатність моделювати складні залежності Погано працюють у шумних умовах або з природними акцентами Потребують ручного проектування ознак (feature engineering) Не масштабуються під великі масиви даних
Моделі на основі глибокого навчання	LSTM/GRU-based генеративні моделі Sequence-to-Sequence (Seq2Seq) Tacotron / FastSpeech (TTS) Transformer / Conformer (ASR, TTS) Whisper, Wav2Vec 2.0, HuBERT (SSL + Seq2Seq)	Навчаються напряму з сирих даних (raw audio або спектрограм) Значно вища точність і природність мовлення Потужне узагальнення навіть на нові голоси / мови Підтримка мультимовності та перенавчання (fine-tuning) Можливість multitask (ASR + Translation + Speaker ID)	Високі обчислювальні вимоги (GPU/TPU) Потреба у великому обсязі навчальних даних (особливо для генерації) Складність у поясненні роботи (чорна скринька) Іноді потребують складної інфраструктури для розгортання

Рекурентні нейронні мережі (RNN). Оскільки мовлення є послідовним процесом, природно використовувати рекурентні нейронні мережі (RNN) для моделювання часових залежностей у мовних даних. RNN здатні враховувати послідовні залежності між елементами вхідного сигналу, що є недоступним для звичайних багатошарових перцептронів. Первинні підходи поєднували RNN із прихованими моделями Маркова (HMM) [7], де RNN виконували локальну класифікацію, а HMM моделювали часову динаміку. Однак така гібридна схема наслідує обмеження HMM, зокрема припущення про незалежність спостережень та вимогу до апріорного знання структури моделі. Це спонукало до впровадження наскрізних архітектур, повністю побудованих на RNN, які виявилися ефективними у завданнях трансдукції послідовностей, таких як автоматичне розпізнавання мовлення.

Прогнозування акустичних параметрів за допомогою DNN. Акустичні характеристики фонем значною мірою залежать від контексту, в якому вони зустрічаються. Ієрархічна структура, яка лежить в основі процесу породження мовлення, зумовлює складні взаємозв'язки між лінгвістичними та акустичними параметрами. Глибокі нейронні мережі (DNN) дозволяють ефективно моделювати ці залежності, забезпечуючи точніше прогнозування акустичних ознак. Порівняно з методами, заснованими на HMM, DNN-архітектури не тільки краще відображають складну лінгвістичну інформацію, але й здатні враховувати довготривалу контекстуальну інформацію, що підвищує якість синтезованого мовлення. До того ж, використання алгоритмів згладжування, таких як MLPG, дозволяє уникнути артефактів, притаманних методам на основі HMM.

Параметричний синтез мовлення відноситься до методу, який використовує технології цифрової обробки сигналів для синтезу мовлення з тексту. У цьому методі він розглядає голосовий процес людини як симуляцію, яка використовує джерело голосового стану для збудження цифрового фільтра, що змінюється в часі, який характеризує резонансні характеристики каналу. Джерелом може бути періодична послідовність імпульсів, яка використовується для представлення вібрації голосових зв'язок голосової мови, або випадковий білий шум для вказівки на невизначену глуху мову. Регулюючи параметри фільтра, він може синтезувати різні типи мовлення [8]. Типові методи включають параметричний синтез голосового органу [9], формантний параметричний синтез [10], синтез мовлення на основі HMM [11] і синтез мовлення на основі глибокої нейронної мережі (DNN) [7, 12].

Вокодери у статистичному синтезі мовлення. Вокодер виступає ключовим компонентом у статистичному параметричному синтезі мовлення, забезпечуючи відтворення мовного сигналу на основі акустичних параметрів. Хоча традиційні системи, зокрема HTS_engine, є простими у реалізації, вони продукують мовлення з низькою природністю. Новітні системи, такі як STRAIGHT, PSOLA, а також моделі, що базуються на синусоїдальних сигналах, забезпечують вищу якість, хоча й поступаються у швидкодії. Сучасні підходи орієнтовані на досягнення балансу між якістю та реалістичністю, наприклад, за допомогою реального часу версій STRAIGHT або використання високоточних вокодерів, таких як WORLD [13].

Двонаправлені рекурентні мережі (BiRNN). Для завдань, які вимагають повного контексту (наприклад, у розпізнаванні мовлення), двонаправлені RNN (BiRNN) є більш придатними, оскільки дозволяють моделювати залежності як з минулого, так і з майбутнього. BiRNN складається з двох RNN, що обробляють послідовність у прямому та зворотному напрямках, після чого об'єднують отримані представлення. Такий підхід дозволяє отримати збагачене представлення вхідного сигналу, що істотно покращує результати при синтезі або аналізі мовлення.

Обмежені машини Больцмана та DBN. Обмежені машини Больцмана (RBM) застосовуються як моделі щільності для представлення спектральних ознак мовлення та часто використовуються для попереднього навчання глибоких нейронних мереж. Глибокі мережі віри (DBN), побудовані на основі RBM, дозволяють спільно моделювати лінгвістичні та акустичні характеристики, що забезпечує більш точне відтворення спектральної огинаючої. Особливу увагу в таких моделях приділяють ефективному кодуванню дискретних (наприклад, V/UV) та неперервних (наприклад, F0) параметрів мовлення.

Глибокий синтез мовлення на основі LSTM. Моделі довготривалої короткочасної пам'яті (LSTM) розв'язують проблему згасання або вибуху градієнта в RNN, що дозволяє зберігати довготривалі залежності в послідовностях. Двонаправлені LSTM застосовуються для точного прогнозування параметрів мовлення, враховуючи як попередній, так і наступний контекст. Такі архітектури забезпечують високу природність синтезованого мовлення.

Процес вивчення репрезентації мовлення має важливе значення для вилучення доречних і практичних характеристик із мовних сигналів, які можна використовувати для виконання різноманітних подальших завдань, таких як ідентифікація мовця, розпізнавання мовлення та розпізнавання емоцій. У той час як традиційні методи розробки інженерних функцій широко

використовувалися, останні досягнення в методах глибокого навчання з використанням навчання під наглядом або без нагляду показали неабиякий потенціал у цій галузі. Незважаючи на це, з'явився новий підхід, заснований на самоконтрольованому навчанні репрезентації, спрямований на розкриття внутрішньої структури мовних даних і отримання репрезентацій, які фіксують базову структуру даних [13]. Цей підхід перевершує традиційні методи розробки функцій і може значно підвищити точність і ефективність в майбутньому. Основна мета цієї нової парадигми полягає в тому, щоб виявити інформативні та значущі характеристики мовних сигналів і удосконалити існуючі підходи. Ми розглянемо різні техніки та архітектури, розроблені протягом багатьох років, у тому числі появу неконтрольованих методів навчання репрезентації, таких як автокодери, генеративні суперницькі мережі (GAN) і системи самоконтрольованого навчання репрезентації. Ми також розглянемо труднощі та обмеження, пов'язані з цими техніками, такі як дефіцит даних, адаптація домену та можливість інтерпретації навчених репрезентацій, дивись таблицю 2.

Таблиця 2

Порівняння моделей за якісними характеристиками

Критерій	Класичні моделі	Глибинне навчання
Точність (WER/TTS MOS)	Низька–середня	Висока
Гнучкість	Низька	Висока
Потреба в даних	Помірна	Висока (особливо для генерації)
Обчислювальні ресурси	Мінімальні	Середні–високі (GPU/TPU)
Мультимовність / Акценти	Слабо підтримується	Добре підтримується
Автоматичне навчання ознак	Ні	Так
Придатність для реального часу	Так (за спрощеної якості)	Так, але за оптимізованих умов (Streaming Transformers, quantization)

Завдяки всебічному аналізу переваг і обмежень різних підходів до навчання репрезентації ми прагнемо зрозуміти, як їх використовувати для підвищення точності систем обробки мовлення.

Під час навчання під контролем модель навчається за допомогою анотованих наборів даних, щоб навчитися зіставляти вхідні дані та вихідні мітки. Набір параметрів, які визначають функцію відображення, оптимізується під час навчання, щоб мінімізувати різницю між прогнозованими та вихідними мітками в навчальних даних. Мета контрольованого навчання полягає в тому, щоб дозволити моделі вивчити корисне представлення або особливості вхідних даних, які можна використовувати для точного прогнозування вихідної мітки для нових, невідомих даних. Наприклад, навчання репрезентації під наглядом під час обробки мовлення за допомогою CNN вивчає особливості мовлення зі спектрограм. CNN можуть ідентифікувати шаблони в спектрограмах, що стосуються розпізнавання мовлення, наприклад ті, що відповідають різним фонемам або словам. На відміну від CNN, які зазвичай вимагають введення спектрограми, RNN можуть безпосередньо отримувати необроблені мовні сигнали як вхідні дані та навчатися виокремлювати ознаки чи представлення, які є релевантними для розпізнавання мовлення або інших завдань обробки мовлення.

Архітектури типу "послідовність до послідовності" (Seq2Seq). Моделі типу Seq2Seq з увагою продемонстрували значний успіх у завданнях перекладу, створення підписів до зображень і розпізнавання мовлення [14]. Їх застосування до задач синтезу мовлення дозволило реалізувати моделі, які прямо перетворюють вхідний текст у спектральні характеристики. Приклади таких систем включають Tacotron, Char2Wav, а також модифікації з прямою або позиційною увагою. Ці архітектури здатні генерувати високоякісне, природне мовлення без необхідності проміжних параметричних перетворень.

Сучасні підходи до синтезу мовлення:

- Wav2Vec 2.0.Розроблена компанією Facebook AI, Wav2Vec 2.0 дозволяє навчати моделі на сирому аудіо з подальшим донавчанням на малій кількості розмічених даних.
- HuBERT. Поєднує підходи кластеризації і masked prediction, дозволяючи ефективно навчатися без учителя. Модель має конкурентну якість у порівнянні з Wav2Vec 2.0.
- Whisper. Модель від OpenAI, навчена на багатомовному корпусі, здатна до транскрибування, перекладу, виявлення мов та інших задач. Відзначається високою точністю на реальних даних.
- Conformer. Поєднує переваги CNN і Transformer у єдиній архітектурі. Покращує якість виявлення локальних і глобальних патернів у мовленні.

У сучасних дослідженнях з автоматичного розпізнавання мовлення (ASR) активно порівнюються різні архітектури глибокого навчання, кожна з яких має свої переваги та обмеження залежно від поставленого завдання. Рекурентні мережі (RNN, LSTM, GRU) демонструють стабільність при роботі з послідовностями, але поступаються у швидкодії трансформерам, які забезпечують високу точність та ефективну паралелізацію завдяки механізмам уваги. Архітектури на основі самонавчання, такі як Wav2Vec 2.0 і HuBERT, дозволяють використовувати неанотовані дані, що є перевагою в умовах обмежених ресурсів, тоді як моделі на кшталт Whisper і Conformer поєднують гнучкість, точність і мультимовну підтримку, водночас залишаючись ресурсомісткими. Порівняльну характеристику основних моделей наведено в таблиці 3, що дозволяє чітко оцінити їхні сильні та слабкі сторони у контексті завдань генерації та розпізнавання мовлення.

Таблиця 3

Порівняння моделей за типами

Модель	Тип	Переваги	Недоліки
RNN/LSTM/GRU	Рекурентні	Простота, стабільність	Обмежена здатність до паралелізації
CNN	Згорткові	Локальна інваріантність	Втрата глобального контексту
Transformer	Увага	Глобальний контекст, паралелізація	Високі обчислювальні витрати
Wav2Vec 2.0	Self-Supervised	Навчання на сирих даних	Потребує великого обсягу даних
HuBERT	Self-Supervised	Сильна генералізація	Складність попередньої обробки
Whisper	Seq2Seq + Attention	Мультимовність, точність	Ресурсоємність
Conformer	Гібридна	Найкраща якість у ASR	Архітектурна складність

Висновки

Проведене дослідження засвідчує, що архітектури глибокого навчання значно підвищили точність обробки мовлення порівняно з традиційними підходами. Рекурентні нейронні мережі (RNN), двонаправлені RNN (BRNN) та моделі довготривалої пам’яті (LSTM) продемонстрували високу здатність до моделювання часових залежностей у мовних сигналах, що критично важливо для задач розпізнавання та синтезу мовлення. Моделі на основі DNN здатні ефективно використовувати контекстну інформацію для прогнозування акустичних характеристик, перевершуючи традиційні HMM у якості результату. Застосування вокодерів нового покоління, таких як STRAIGHT, WORLD та інші, дозволяє досягати вищої природності

синтезованого мовлення, хоча проблема узгодження якості та швидкодії все ще залишається актуальною. Розробки на основі seq2seq архітектур з механізмом уваги, зокрема моделі типу Tacotron, відкривають нові можливості для побудови більш адаптивних і точних систем синтезу мовлення, здатних моделювати мел-спектрограми без необхідності використання ручного маркування або складного препроцесингу. Попри значні досягнення, низка викликів залишається нерозв'язаною. Серед них – нестабільність моделей під час навчання, проблема надмірної гладкості спектральних ознак, складність роботи з великими мовними корпусами та потреба в адаптації до індивідуальних голосів. Подальші дослідження мають бути спрямовані на вдосконалення моделей, здатних поєднувати високу якість, адаптивність та реальний час обробки. Отже, глибоке навчання формує потужний теоретичний та практичний фундамент для розвитку мовленнєвих технологій нового покоління, забезпечуючи нові можливості в таких сферах, як голосові помічники, автоматичне субтитрування, інтерфейси людина-машина та інклюзивні технології для осіб із порушеннями мовлення.

Список використаної літератури:

1. Self-Supervised Speech Representation Learning: A Review / A. Mohamed et al. *IEEE Journal of Selected Topics in Signal Processing*. 2022. P. 1–34. URL: https://doi.org/10.1109/jstsp.2022.3207050.

2. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units / W.-N. Hsu et al. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021. Vol. 29. P. 3451–3460. URL: https://doi.org/10.1109/taslp.2021.3122291

3. Hastad J., Risse K. On Bounded Depth Proofs for Tseitin Formulas on the Grid; Revisited. *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, Denver, CO, USA, 31 October – 3 November 2022. 2022. URL: https://doi.org/10.1109/focs54457.2022.00110

4. Conformer: Convolution-augmented Transformer for Speech Recognition / A. Gulati et al. *Interspeech 2020*. ISCA, 2020. URL: https://doi.org/10.21437/interspeech.2020-3015.

5. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions / J. Shen et al. *ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 15–20 April 2018. 2018. URL: https://doi.org/10.1109/icassp.2018.8461368.

6. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations / J. Giorgi et al. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.acl-long.72.

7. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis / T. Yoshimura et al. *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*. ISCA, 1999. URL: https://doi.org/10.21437/eurospeech.1999-513.

8. Xu S.H. Study on HMM-Based Chinese Speech Synthesis. Beijing : Beijing University of Posts and Telecommunications, 2007.

9. Sotelo J., Mehri S., Kumar K., Santos J.F., Kastner K., Courville A., Bengio Y. Char2wav: End-to-end Speech Synthesis // Proceedings of the International Conference on Learning Representations Workshop, Toulon, France, 24–26 April 2017.

10. Klatt D. H. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*. 1980. Vol. 67, no. 3. P. 971–995. URL: https://doi.org/10.1121/1.383940.

11. Moulines E., Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*. 1990. Vol. 9, no. 5-6. P. 453–467. URL: https://doi.org/10.1016/0167-6393(90)90021-z.

12. Ze H., Senior A., Schuster M. Statistical parametric speech synthesis using deep neural networks. *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 26–31 May 2013. 2013. URL: https://doi.org/10.1109/icassp.2013.6639215.

13. Morise M., Yokomori F., Ozawa K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*. 2016. E99.D, no. 7. P. 1877–1884. URL: https://doi.org/10.1587/transinf.2015edp7457.

14. Luong T., Pham H., Manning C. D. Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Stroudsburg, PA, USA, 2015. URL: https://doi.org/10.18653/v1/d15-1166.

Автори статті

Іщеряков Сергій – кандидат технічних наук, доцент, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.

ORCID: 0009-0007-5961-8218

Попов Антон – аспірант, Державний університет інформаційно-комунікаційних технологій, Київ, Україна.

ORCID: 0009-0006-7557-094X

Authors of the article

Ishcheriakov Serhii – Candidate of Sciences (technical), Associate Professor, State University of Information and Communication Technologies, Kyiv, Ukraine.

ORCID: 0009-0007-5961-8218

Popov Anton – postgraduate, State University of Information and Communication Technologies, Kyiv, Ukraine.

ORCID: 0009-0006-7557-094X