УДК 004.032.26:004.8
DOI: 10.31866/2617-796X.5.1.2022.261297

Ткаченко Костянтин,
кандидат економічних наук,
доцент кафедри інформаційних технологій та дизайну,
Державний університет інфраструктури та технологій,
Київ, Україна
tkachenko.kostyantyn@gmail.com
https://orcid.org/0000-0003-0549-3396

Брусєнцев Владислав,
магістрант, кафедра інформаційних технологій та дизайну,
Державний університет інфраструктури та технологій,
Київ, Україна
vladbrusentcev1@gmail.com
https://orcid.org/0000-0002-8106-5855

ВИКОРИСТАННЯ НЕЙРОННИХ МЕРЕЖ ПІД ЧАС РОЗПІЗНАВАННЯ ГОЛОСОВИХ КОМАНД

Метою статті є дослідження, аналіз і розгляд загальних проблем та перспектив щодо розробки систем розпізнання голосових команд з використанням можливостей нейронних мереж та новітніх нейромережевих технологій.

Методами дослідження є методи семантичного аналізу основних понять цієї предметної сфери (системи розпізнавання голосових команд). У статті розглянуто наявні системи й алгоритми розпізнання.

Новизною проведеного дослідження є аналіз функціонування сучасних систем розпізнавання голосових команд, результати якого можуть застосовуватися під час розробки власної системи розпізнавання на основі використання покращених мовленнєвих моделей і рекурентної нейронної мережі, що навчається.

Висновки. Доведено ефективність використання нейронних мереж для завдань розпізнавання голосових команд. Розроблено систему розпізнавання мовлення на основі нейронних мереж з використанням покращеної мовленнєвої моделі.

Ключові слова: нейронні мережі; навчання нейронних мереж; рекурентні нейронні мережі; розпізнання; штучний інтелект; голосові команди.

Вступ. На сьогодні в теорії розпізнавання образів і технологій, що її підтримують, усе більше уваги приділяють системам розпізнавання голосових команд. Різні категорії користувачів широко використовують такі системи, що відкривають нові можливості. Наприклад, у поліклініці лікар може вимовляти діагнози, які відразу заносять в електронну картку, а в системах інтернету речей (IoT) можна керувати автомобілем чи домашніми пристроями, надаючи їм голосові команди (https://iotukraine.com).

Системи розпізнавання голосових команд дають змогу, зокрема, безпечно працювати користувачам; голосового введення інформації (тексту, команд тощо); голосового пошуку інформації; синтезу речень.

Успіхи в розпізнаванні та синтезі мовлення сприяли появі голосових помічників, які можуть спілкуватися з користувачами та виконувати різні голосові команди. Такий спосіб комунікації підвищує продуктивність виробництва, ефективність навчання, комфортність роботи із системами електронної комерції. Усе це досягається за допомогою інтелектуалізації процесів розпізнавання образів, використовуючи, зокрема, у цих процесах нейронні мережі та нейромережеві технології (Sokolov and Savchenko, 2019; Субботін, Олійник, А. та Олійник, О., 2009 ; Zheng, Meng and Jin, 2011).

Нейронну мережу можна використовувати як інструмент інтелектуалізації систем розпізнавання голосових команд (Система розпізнавання голосу; Квитко, 2016; Hinton et el., 2012). У системі розпізнавання голосу, наприклад, розпізнавання голосових команд забезпечується відповідною системою розпізнавання, гарантуючи безпечне використання мультимедіа під час керування автомобілем. Але через наявність технологічних обмежень не всі голосові команди розпізнаються, тому система обробляє лише обмежену кількість таких голосових команд керування.

Нейронні мережі можуть виконувати різні завдання, включаючи розпізнавання голосу, і мають деякі переваги перед традиційними методами, такими як (Крюкова, 2018; Амосов, Иванов и Жиганов, 2017; Робейко та Мартиненко, 2014):

– приховані моделі Маркова;
– «розсувне вікно»;
– моделі-«заповнювачі».

За допомогою нейронних мереж можна реалізувати адаптивні системи з постійними можливостями навчання. Завдяки цьому нейронні мережі набувають високої популярності, бо можуть «навчитися» розв’язувати будь-яке завдання.

Використання системи розпізнавання голосу на виробництві підвищить продуктивність і покращить взаємодію користувача із системою.

На сьогодні є багато робіт, присвячених проблемам розпізнавання голосових команд, в яких розглядають:

– методи навчання голосового управління (Muda, Mumtaj and Elamvazuthi, 2010);

– алгоритми машинного навчання для перетворення інформації з текстового представлення в голосове (Li, 2021; Bengio, 2009);

– методи перетворення голосу (Ault et el., 2018; Gers, Schraudolph and Schmidhuber, 2002).

На сьогодні є багато систем розпізнавання мовлення, таких як Microsoft API, Google API та CMU Sphinx (Toda, Nakagiri and Shikano, 2012; Desai et el., 2010).

Зокрема, у Microsoft API, Google API (Toda, Nakagiri and Shikano, 2012) акцент на розпізнаванні мовлення з відкритим кодом у різних середовищах – це аудіозаписи, відібрані з різних джерел, з розрахунком коефіцієнта помилок у словах.

У CMU Sphinx (Desai et el., 2010; Këpuska and Bohouta, 2017) розглянуто системи з відкритим кодом (CMU Sphinx, Kaldi, Julius, HTK, iAtros, RWTH ASR і Simon) та системи із закритим кодом (Dragon, Mobile SDK, Google Speech Recognition API, Siri, Yandex SpeechKit і Microsoft Speech API).

Усі перераховані вище системи використовують приховану модель Маркова; потребують обширних словників фонем та їх станів для організації коректної роботи системи щодо розпізнавання. Тому актуальність проблеми розробки нової системи розпізнання мовлення, що буде заснована на нейронних мережах і зможе використати їх переваги для збільшення точності й швидкості розпізнання, не викликає сумнівів.

Метою дослідження є аналіз і розгляд загальних проблем та перспектив щодо розробки систем розпізнання голосових команд з використанням можливостей нейронних мереж та новітніх нейромережевих технологій.

Результати дослідження. Ефективність систем розпізнання голосових команд можна підвищити, покращивши властивості відповідних мовленнєвих моделей (Модель мовленнєвої комунікації). Мовленнєва модель містить реалізацію діалогу людини в конкретній ситуації мовленнєвого спілкування. *Модель мовлення* – це ціле сімейство звуків, іноді дуже різних за складом векторних знаків.

Біфони, які описують фонему в поєднанні з попередньою або наступною фонемою, використовуються для опису початку або кінця фрагмента мови, а також коли недостатньо даних для побудови станів *трифону*. Контекстуальні фонеми називаються *монофонами*.

Для підвищення специфічності мовленнєвих моделей можна використовувати такі відомі підходи:

– збільшення кількості фонем, наприклад, за допомогою розщеплення фонем і самостійної їх обробки в словнику;

– збільшення кількості станів у фоновому режимі (наприклад, від одного до трьох станів для однієї фонеми);

– залежність від контексту фонем, наприклад використання трифонів (усі розглянуті комбінації фонем з попередніми та наступними звуками як окремими акустичними об’єктами, для яких потрібно будувати свої стани);

– моделювання варіацій вимови слів, наприклад включення кількох варіантів вимови слова до словника.

Оптимізація рівня специфічності мовленнєвих моделей певної бази даних є трудомістким процесом. Це не стосується конкретно нейронних мереж. Топологія фонеми під час навчання зазвичай незначна, але іноді використовується до трьох станів на фонему з простим переходом зліва направо.

В українській мові кількість фонем становить близько 50 (кількість не фіксована – кількість загальних біфонів або трифонів можна заздалегідь зарахувати до окремих фонем).

У табл. 1 наведено приклад залежності точності системи від кількості станів на фонему. З табл. 1 видно, що було взято лише п'ять епох, тому що одна епоха призводить до недообладнання відповідної компоненти (системи, нейронних мереж, словника тощо), а надлишок епох – до переобладнання. Зі збільшенням кількості епох ваги нейронної мережі змінюються все більше і більше разів.

Таблиця 1

Залежність точності системи від кількості станів фонем

Епоха	Точність слів, %		
	1 стан на фонему	1... 3 стани на фонему	3 стани на фонему
1	81,9	85,1	85,0
2	85,8	86,4	86,2
3	86,3	86,9	87,0
4	86,3	87,0	88,0
5	86,1	86,9	88,0

Аналізуючи дані з табл. 1, можна сказати, що зі збільшенням ітерацій більша точність слова була при точно трьох станах на фонему (88,5 %), а найменша – при точно одному стані на фонему.

Після аналізу табл. 1 для ситуації «1... 3 стани на фонему» можна побачити, що спочатку точність вища, а в кінцевому підсумку точність ітерацій усе ще залишається трохи вищою, ніж з одним станом на фонему.

Отже, можна зробити висновок, що найкращі результати отримали, коли додаткові умови на фонему були корисними та навченими належним чином. Тому можемо однозначно стверджувати, що в міру збільшення станів точність словникових слів зростає.

Гнучкість словникового запасу також впливає на покращення результатів роботи систем розпізнавання голосових команд, де кожне слово має кілька варіантів вимови. Це дає змогу забезпечити більшу точність розпізнавання фонем, що є в інформаційній базі системи.

Використовуючи таку технологію, можна побачити, що деякі слова мають дуже схожу вимову, тому система часто може робити помилки під час розпізнавання цих слів.

Штучні нейронні мережі, що використовуються в системах розпізнавання звукових образів, засновані на структурі простих нелінійних обчислювальних елементів. Серед таких нейронних мереж виділимо рекурентні нейронні мережі. RNN (Recurrent Neural Networks) – клас штучних нейронних мереж, зв'язки між вузлами яких утворюють орієнтований у часі граф. Це створює внутрішній стан мережі, дозволяючи їй проявляти з часом динамічну поведінку (Ali, Eltayeb and Abusail, 2017).

Можна стверджувати, що використання глибоких нейронних мереж і рекурентних нейронних мереж є однією з найбільш сучасних технологій автоматичного розпізнавання мови.

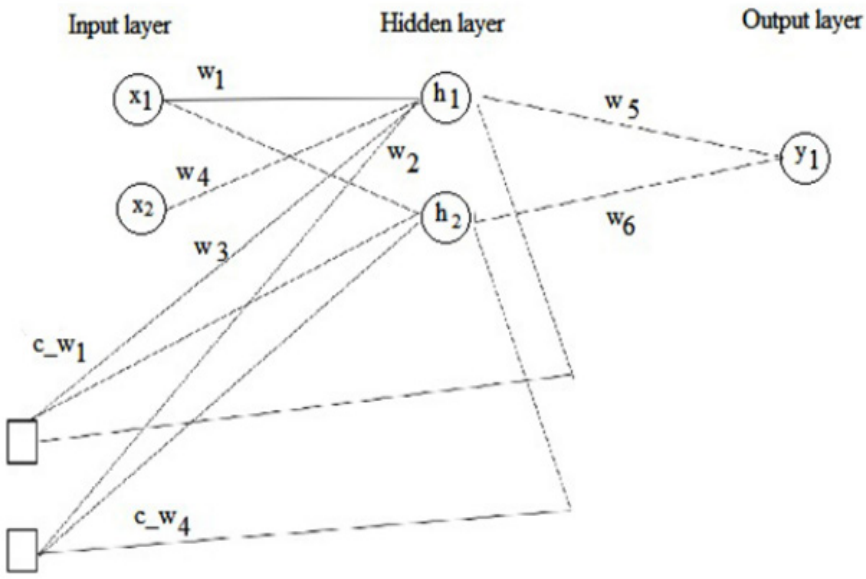

Рис. 1. Вузли нейронної мережі

Вузол (рис. 1) має N входів, позначених x_1, x_2, …, x_n, які підсумовуються з вагами w_1, w_2,…, w_n. Спочатку інформація надсилається з вхідного рівня (шару) за ваговими значеннями до прихованого:

$$h_1 = (x_1 * w_1) + (x_2 * w_4),$$

$$h_2 = (x_1 * w_2) + (x_2 * w_3),$$

де h – прихований прошарок.

Потім інформація надсилається від прихованих нейронів на рівень (шар) затримки часу та на вихід мережі:

$$c_1 = h_1 , c_2 = h_2;$$

$$y_1 = (h_1 * w_5) + (h_2 * w_6).$$

Далі дані записуються на рівень (шар) затримки часу, а потім сигнал «запускається» знову, додаються тільки сигнали від шару затримки:

$$h_1 = (x_1 * w_1) + (x_2 * w_4) + (c_1 * c\_w_1) + (c_2 * c\_w_3).$$

І на другому прихованому нейроні:

$$h_2 = (x_1 * w_2) + (x_2 * w_3) + (c_1 * c\_w_2) + (c_2 * c\_w_4).$$

Потім отримані дані знову надсилаються на рівень (шар) затримки та виводяться:

$$c_1 = h_1 , c_2 = h_2;$$

$$y_1 = (h_1 * w_5) + (h_2 * w_6).$$

Нейрон має ваги, які помножуються на отримані дані, що призводить до зміненої реакції на вхідний сигнал.

Потім модифіковані реакції в нейроні підсумовуються і переходять (як вхідні дані) до функції активації.

Функція активації формує (складає) відповідь з отриманої суми. Можна використовувати порогову функцію або сигмоїдальну (наприклад, гіперболічний тангенс і логістичну функцію).

Застосування порогової функції передбачає таке: коли є результат підсумовування і деякий поріг, то їх необхідно порівняти. Якщо загальний результат перевищує поріг, то нейрон виведе як результат значення 1, а якщо результат не перевищує поріг – значення 0.

Гіперболічний тангенс перетворює загальний результат у число від -1 до 1. Для цього скористайтеся формулою (Beck, 2018):

$$exp(out) - exp(-out) / exp(out) + exp(-out),$$

де exp – експоненціальна функція.

Логістична функція перетворює загальний результат у число, що належить відрізку [0, 1]. Для цього слід скористатися формулою (Beck, 2018):

$$1 / (1 + exp(-out)).$$

Зрештою, виявляється, що рекурентні нейронні мережі здатні до короткочасної пам'яті. Для розв'язання проблеми класифікації мовлення можна використовувати багато інструментів.

Для реалізації системи розпізнавання голосових команд і розробки її програмного забезпечення мовою програмування обрано Python (Millstein, 2018). Python – високорівнева об'єктно-орієнтована мова програмування із сильною динамічною типізацією. Ця мова програмування має такі переваги: зручний інструмент для розв'язування математичних задач; нескладний синтаксис; наявність великої кількості сторонніх бібліотек, які можуть застосовуватися під час написання програмного коду; відкритий код.

Для розв'язання задачі класифікації мовлення на першому етапі слід вибрати бібліотеку з навчальними даними (для навчання відповідної нейронної мережі). Для того щоб нейронна мережа забезпечувала високу точність відображення голосової команди, необхідно мати досить велику базу даних, яка буде містити навчальні послідовності.

Для розв'язання задачі класифікації мовлення в роботі використано бібліотеку TIMIT. Ця бібліотека розроблена з метою надання мовленнєвих даних для набуття акустично-фонетичних знань, а також для розробки й оцінки систем автоматичного розпізнавання мовлення.

TIMIT – бібліотека, розроблена для дослідницьких проєктів у галузі оборони й управління інформаційними науками та технологіями (Glackin et el., 2018).

Фонеми оцифровані на частоті 20 кГц з фільтром налаштування 10 кГц. Після цього фонеми були відфільтровані та зменшені до 16 кГц.

Кожна з фонем має мітку. Ці мітки представляють дещо проміжний рівень інформації між фонематичним і акустичним.

Для навчання мережі використовувався фреймворк RNNLM (рекурентні нейронні мережі) (Lipeika, Lipeikienė and Telksnys, 2002).

На відміну від нейронних мереж прямого зв'язку, RNN можуть використовувати свою внутрішню пам'ять для обробки довільних послідовностей вхідних даних. Це дає змогу застосовувати їх до таких завдань, як безсегментне безперервне розпізнавання рукописного тексту та розпізнавання мовлення (рис. 2).

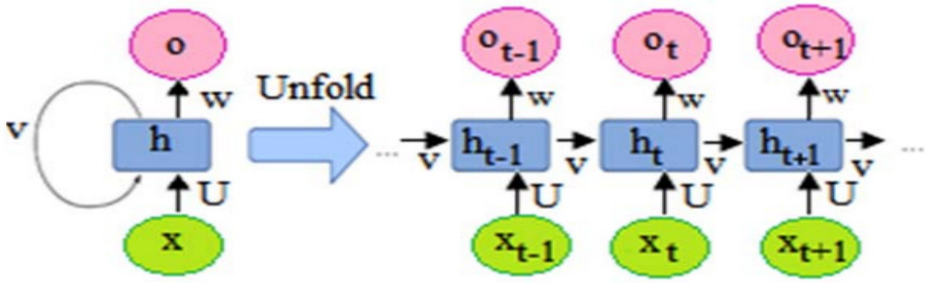

Рис. 2. Архітектура моделі рекурентної нейронної мережі

Рекурентна модель нейронної мережі (рис. 2), вхідний рівень якої використовує 1-із-N подання попереднього слова, яке поєднується з попереднім станом прихованого шару s(t) за допомогою сигмоїдальної функції активації. Вихідний шар y(t) має такий же розмір, як w(t). Навчання відбуватиметься за алгоритмом стохастичного градієнта.

Отримання вагових коефіцієнтів у нейронній мережі можна моделювати як нелінійну глобальну оптимізаційну задачу. Об'єктивну функцію для оцінки придатності або помилки певного вагового вектора можна сформувати таким чином:

1. Вага в мережі встановлюється відповідно до вектора ваги.

2. Мережа оцінюється за послідовністю навчання. Як правило, сума квадратів різниць між прогнозами та цільовими значеннями, заданими в навчальній послідовності, використовується для представлення помилки в поточному векторі ваги.

3. Для мінімізації цієї цільової функції можна застосувати довільні методи глобальної оптимізації.

Запропонована система має аналоги, які реалізовані за схожим принципом і можуть виконувати подібні завдання. Однак ця система реалізована з використанням максимальної гнучкості, щоб її можна було легко адаптувати до поставлених завдань. У порівнянні із системою розпізнавання мовлення, яку надає Google, запропонована система не вимагає постійного підключення до інтернету.

Автономна робота системи розпізнавання голосових команд має перевагу через зменшення вимог до роботи з нею. Система розпізнавання мовлення завдяки постійному підключенню до інтернету використовує хмарні технології, що позитивно позначається на швидкості роботи системи, простоті адаптації користувачів

до роботи з нею; на переналаштування системи потрібно в декілька разів менше часу, оскільки вона не вимагає повторного навчання відповідної нейромережі.

Слід також зазначити, що система від Google має можливість постійно навчатися, оновлюючи базу даних (з навчальними послідовностями). Маючи великі потужності та використовуючи хмарні технології, система може навчатися в режимі реального часу та легко адаптуватися до нового навантаження.

Запропонована система має лише один прихований шар, що достатньо для розв'язання невеликого кола завдань цієї системи. Збільшення кількості прихованих шарів також збільшує час, необхідний для навчання відповідної нейронної мережі. До того ж система використовує потужність хмарних технологій і переналаштування системи відбувається миттєво. Точність розробленої системи досягла 84,4 % при навчанні системи за 10 ітерацій. Система розпізнавання мовлення показує результат з точністю до 98 % (Swamy and Ramakrishnan, 2013).

Використання хмарних технологій підвищує точність і швидкість роботи системи, але накладає на неї певні обмеження щодо сфер використання.

У порівнянні із системою PocketSphinx, яка також є автономною, але не використовує хмарні технології, що притаманні системі розпізнання Speech, запропонована система показала найкращий результат. Точність системи PocketSphinx досягла лише 79,2 % при подібних параметрах під час навчання систем (Huggins-Daines et el., 2006). Насамперед це пов'язано з використанням різних топологій нейронних мереж.

Слід також зазначити, що системи великих корпорацій (наприклад, Microsoft, Google і Yandex) вимагають ліцензування, а весь код не доступний для пересічного розробника програмного забезпечення.

Хоча бібліотека Kaldi дає змогу використовувати систему за ліцензією Apache 2.0, вона майже не накладає на неї обмежень (Cutajar et el., 2013).

Запропонована система показала оптимальні результати точності під час використання бібліотеки Kaldi. Після експерименту отримали відповідні результати (табл. 2).

Таблиця 2

Результати порівняння точності та швидкості

Система	Розроблена система	PocketSphinx	Розпізнання мовлення Google
Точність розпізнання, %	84,4 %	79,2 %	98 %
Швидкість розпізнання	0,6	0,5… 1	0,5
Використаний алгоритм	Нейронна мережа	Прихована модель Маркова	Прихована модель Маркова
Мова програмування	Python	C/Java	C

Висновки. Проведене дослідження підтвердило актуальність заявленої тематики роботи, а отримані результати використано під час проєктування та розробки системи розпізнавання мовлення на основі нейронних мереж із застосуванням покращеної мовленнєвої моделі.

Мовою програмування обрано Python, для реалізації використано фреймворк Kaldi.

Навчання системи відбувалося на даних бібліотеки TIMIT. У ролі архітектури обрано рекурентну нейронну мережу.

Функціонування розробленої системи порівняли з розпізнаванням мовлення в системах від Google і PocketSphinx.

Результати проведеного порівняння підтвердили ефективність використання нейронних мереж для розпізнання голосових команд. Упроваджена система досягла точності 84,4 % при швидкості розпізнання 0,6 с.

СПИСОК ПОСИЛАНЬ

Амосов, О.С., Иванов, Ю.С. и Жиганов, С.В., 2017. Локализация человека в кадре видеопотока с использованием алгоритма на основе растущего нейронного газа и нечёткого вывода. *Компьютерная оптика*, [online] 41 (1), с.46-58. Доступно: <https://doi.org/10.18287/2412-6179-2017-41-1-46-58> [Дата звернення 18 квітня 2022].

Інтернет речей. [online] Доступно: <https://iotukraine.com> [Дата звернення 25 квітня 2022].

Квитко, М.В., 2016. Распознавание речи с помощью глубоких рекуррентных нейронных сетей. В: *System Analysis and Information Technologies 18-th International Conference SAIT 2016*. Kyiv, Ukraine, [online] 30 May-2 June 2016, pp.223-224. Kyiv: Kyiv Polytechnic Institute. Доступно: <http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8- 66f11e-e074cd/sait2016ebook.pdf> [Дата звернення 18 квітня 2022].

Крюкова, Г., 2018. Приховані моделі Маркова: регуляризація та застосування в прикладних задачах. В: *Сучасні проблеми математики та її застосування в природничих науках і інформаційних технологіях.* Міжнародна наукова конференція. Чернівці, Україна, [online] 17-19 вересня 2018 р., с.147. Доступно: <http://ekmair.ukma.edu.ua/handle/123456789/15604> [Дата звернення 21 квітня 2022].

Модель мовленнєвої комунікації. *Навчальні матеріали онлайн.* [online] Доступно: <https://pidru4niki.com/12810419/psihologiya/model_movlennyevoyi_komunikatsiyi> [Дата звернення 18 квітня 2022].

Робейко, В. та Мартиненко, М., 2014. Моделювання звуків-заповнювачів і розтягненої вимови звуків у словах у системі автоматичного розпізнавання українського спонтанного мовлення. In: *Ukrainica VI. Současná ukrajinistika. Problémy jazyka, literatury a kultury.* Sborník vědeckých článků z mezinárodní konference «VI Olomoucké sympozium ukrajinistů střední a východní Evropy». Olomouc, Česko, 21-23.08.2014. Olomouc: Univerzita Palackeho v Olomouci, с.424-427.

Система розпізнавання голосу. *Kia*. [online] Доступно: <http://webmanual.kia.com/STD_GEN5W_8/AVNT/EU/Ukrainian/voicerecognitionsystem.html> [Дата звернення 15 квітня 2022].

Субботін, С.О., Олійник, А.О. та Олійник, О.О., 2009. *Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей.* Запоріжжя: ЗНТУ.

Ahmad, M.A., Baker, J.H., Tvoroshenko, I. and Lyashenko, V. 2019. Computational Complexity of the Accessory Function Setting Mechanism in Fuzzy Intellectual Systems. *International Journal*

of Advanced Trends in Computer Science and Engineering, [online] 8 (5), pp.2370-2377. Available at: <https://doi.org/10.30534/ijatcse/2019/77852019> [Accessed 21 April 2022].
Ali, A.T., Eltayeb, E.B. and Abusail, E.A.A., 2017. Voice Recognition Based Smart Home Control System. *International Journal of Engineering Inventions*, 6 (4). pp.1-5.
Ault, S.V., Perez, R.J., Kimble, C.A. and Wang J. 2018. On Speech Recognition Algorithms. *International Journal of Machine Learning and Computing*, [online] 8 (6). pp.518-523. Available at: <https://doi.org/10.18178/ijmlc.2018.8.6.739> [Accessed 21 April 2022].
Beck, M.W., 2018. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software*, [online] 85 (11). pp.1-20. Available at: <https://doi.org/10.18637/jss.v085.i11> [Accessed 21 April 2022].
Bengio, Y., 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, 2 (1). pp.1-127.
Cutajar, M., Gatt, E., Grech, I., Casha, O. and Micallef, J., 2013. Comparative study of automatic speech recognition techniques. *IET Signal Processing*, [online] 7 (1), pp.25-46. Available at: <https://doi.org/10.1049/iet-spr. 2012.0151> [Accessed 23 April 2022].
Desai, S., Black, A.W., Yegnanarayana, B. and Prahallad, K., 2010. Spectral Mapping Using Artificial Neural Networks for Voice Conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, [online] 18 (5), pp.954-964. Available at: <https://doi.org/10.1109/TASL.2010.2047683> [Accessed 23 April 2022].
Gers, F., Schraudolph, N. and Schmidhuber, J., 2002. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, 3, pp.115-143.
Glackin, C., Wall, J., Chollet, G., Dugan, N. and Cannings, N., 2018. TIMIT and NTIMIT Phone Recognition Using Convolutional Neural Networks. In: *Pattern Recognition Applications and Methods*. 7th International Conference, ICPRAM 2018, Funchal, Madeira, Portugal, [online] 16-18 January 2018. Revised Selected Papers, pp.89-100. Available at: <https://doi.org/10.1007/978-3-030-05499-1_5> [Accessed 21 April 2022].
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, Abdel-rahman, Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29 (6), pp.82-97.
Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M. and Rudnicky, A.I., 2006. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, [online] 14-19 May 2006. Available at: <https://doi.org/10.1109/ICASSP.2006.1659988> [Accessed 23 April 2022].
Këpuska, V. and Bohouta, G., 2017. Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Journal of Engineering Research and Application*, [online] 7 (3), pp.20-24. Available at: <https://doi.org/10.9790/9622-0703022024> [Accessed 23 April 2022].
Li, N., 2021. An improved machine learning algorithm for text-voice conversion of English letters into phonemes. *Journal of Intelligent & Fuzzy Systems*, [online] 40 (2), pp.2743-2753. Available at: <https://doi.org/10.3233/JIFS-189316> [Accessed 21 April 2022].
Lipeika, A., Lipeikienė, J. and Telksnys, L., 2002. Development of Isolated Word Speech Recognition System. *Informatica*, [online] 13 (1), pp.37-46. Available at: <https://doi.org/10.3233/INF-2002-13103> [Accessed 19 April 2022].
Millstein, F., 2018. *Natural Language Processing With Python: Natural Language Processing Using NLTK.* Create Space Independent Publishing Platform.

Muda, L., Mumtaj, B. and Elamvazuthi, I., 2010. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing*, 2 (3), pp.138-143.
Sokolov, A. and Savchenko, A.V., 2019. Voice command recognition in intelligent systems using deep neural networks. In: *2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. Herlany, Slovakia, [online] 24-26 January 2019, pp.113-116. IEEE. Available at: <> [Accessed 21 April 2022].
Swamy, S., and Ramakrishnan, K.V., 2013. An efficient speech recognition system. *Computer Science & Engineering: An International Journal (CSEIJ)*, [online] 3 (4), pp.21-27. Available at: <https://doi.org/10.5121/cseij.2013.3403> [Accessed 21 April 2022].
Toda, T., Nakagiri, M. and Shikano, K., 2012. Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, [online] 20 (9), pp.2505-2517. Available at: <https://doi.org/10.1109/TASL.2012.2205241> [Accessed 22 April 2022].
Zheng, Y., Meng, Y. and Jin, Y., 2011. Object Recognition using Neural Networks with Bottom-up and Top-down Pathways. *Neurocomputing*, 74, pp.3158-3169.

REFERENCES

Ahmad, M.A., Baker, J.H., Tvoroshenko, I. and Lyashenko, V. 2019. Computational Complexity of the Accessory Function Setting Mechanism in Fuzzy Intellectual Systems. International Journal of Advanced Trends in Computer Science and Engineering, [online] 8 (5), pp.2370-2377. Available at: <https://doi.org/10.30534/ijatcse/2019/77852019> [Accessed 21 April 2022].
Ali, A.T., Eltayeb, E.B. and Abusail, E.A.A., 2017. Voice Recognition Based Smart Home Control System. International Journal of Engineering Inventions, 6 (4). pp.1-5.
Amosov, O.S., Ivanov, Iu.S. and Zhiganov, S.V., 2017. Lokalizatciia cheloveka v kadre videopotoka s ispolzovaniem algoritma na osnove rastushchego neironnogo gaza i nechetkogo vyvoda [Localization of a person in the frame of a video stream using an algorithm based on growing neural gas and fuzzy inference]. *Kompiuternaia optika*, [online] 41 (1), pp.46-58. Available at: <https://doi.org/10.18287/ 2412-6179-2017-41-1-46-58> [Accessed 18 April 2022].
Ault, S.V., Perez, R.J., Kimble, C.A. and Wang J. 2018. On Speech Recognition Algorithms. International Journal of Machine Learning and Computing, [online] 8 (6). pp.518-523. Available at: <https://doi.org/10.18178/ijmlc.2018.8.6.739> [Accessed 21 April 2022].
Beck, M.W., 2018. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software,* [online] 85 (11). pp.1-20. Available at: <https://doi.org/10.18637/jss.v085.i11> [Accessed 21 April 2022].
Bengio, Y., 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning,* 2 (1). pp.1-127.
Cutajar, M., Gatt, E., Grech, I., Casha, O. and Micallef, J., 2013. Comparative study of automatic speech recognition techniques. *IET Signal Processing,* [online] 7 (1), pp.25-46. Available at: <https://doi.org/10.1049/iet-spr. 2012.0151> [Accessed 23 April 2022].
Desai, S., Black, A.W., Yegnanarayana, B. and Prahallad, K., 2010. Spectral Mapping Using Artificial Neural Networks for Voice Conversion. *IEEE Transactions on Audio, Speech, and Language Processing,* [online] 18 (5), pp.954-964. Available at: <https://doi.org/10.1109/TASL.2010.2047683> [Accessed 23 April 2022].

Gers, F., Schraudolph, N. and Schmidhuber, J., 2002. Learning precise timing with LSTM recurrent networks. *Journal of Machine Learning Research*, 3, pp.115-143.
Glackin, C., Wall, J., Chollet, G., Dugan, N. and Cannings, N., 2018. TIMIT and NTIMIT Phone Recognition Using Convolutional Neural Networks. In: *Pattern Recognition Applications and Methods. 7th International Conference, ICPRAM 2018*, Funchal, Madeira, Portugal, [online] 16-18 January 2018. Revised Selected Papers, pp.89-100. Available at: <https://doi.org/10.1007/978-3-030-05499-1_5> [Accessed 21 April 2022].
Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, Abdel-rahman, Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29 (6), pp.82-97.
Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M. and Rudnicky, A.I., 2006. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, [online] 14-19 May 2006. Available at: <https://doi.org/10.1109/ICASSP.2006.1659988> [Accessed 23 April 2022].
Këpuska, V. and Bohouta, G., 2017. Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Journal of Engineering Research and Application,* [online] 7 (3), pp.20-24. Available at: <https://doi.org/10.9790/9622-0703022024> [Accessed 23 April 2022].
Kriukova, H., 2018. Prykhovani modeli Markova: rehuliaryzatsiia ta zastosuvannia v prykladnykh zadachakh [Hidden Markov models: regularization and application in applied problems]. In: *Suchasni problemy matematyky ta yii zastosuvannia v pryrodnychykh naukakh i informatsiinykh tekhnolohiiakh* [Modern problems of mathematics and its application in natural sciences and information technologies]. International scientific conference. Chernivtsi, Ukraine, [online] 17-19 September 2018, p.147. Available at: <http://ekmair.ukma.edu.ua/handle/123456789/15604> [Accessed 21 April 2022].
Kvitko, M.V., 2016. Raspoznavanie rechi s pomoshchiu glubokikh rekurrentnykh neironnykh setei [Speech recognition using deep recurrent neural networks]. In: *System Analysis and Information Technologies 18-th International Conference SAIT 2016*. Kyiv, Ukraine, [online] 30 May-2 June 2016, pp.223-224. Kyiv: Kyiv Polytechnic Institute. Available at: <http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8- 66f11ee074cd/sait2016ebook.pdf> [Accessed 18 April 2022].
Li, N., 2021. An improved machine learning algorithm for text-voice conversion of English letters into phonemes. *Journal of Intelligent & Fuzzy Systems,* [online] 40 (2), pp.2743-2753. Available at: <https://doi.org/10.3233/JIFS-189316> [Accessed 21 April 2022].
Lipeika, A., Lipeikienė, J. and Telksnys, L., 2002. Development of Isolated Word Speech Recognition System. Informatica, [online] 13 (1), pp.37-46. Available at: <https://doi.org/10.3233/INF-2002-13103> [Accessed 19 April 2022].
Millstein, F., 2018. Natural Language Processing With Python: Natural Language Processing Using NLTK. Create Space Independent Publishing Platform.
Model movlennievoi komunikatsii [Model of speech communication]. *Navchalni materialy onlain*. [online] Available at: <https://pidru4niki.com/12810419/psihologiya/model_movlennyevoyi_komunikatsiyi> [Accessed 18 April 2022].
Muda, L., Mumtaj, B. and Elamvazuthi, I., 2010. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing,* 2 (3), pp.138-143.
Robeiko, V. and Martynenko, M., 2014. Modeliuvannia zvukiv-zapovniuvachiv i roztiahnenoi vymovy zvukiv u slovakh u systemi avtomatychnoho rozpiznavannia ukrainskoho spontannoho

movlennia [Modeling of filler sounds and stretched pronunciation of sounds in words in the system of automatic recognition of Ukrainian spontaneous speech]. In: *Ukrainica VI. Současná ukrajinistika. Problémy jazyka, literatury a kultury. Sborník vědeckých článků z mezinárodní konference* ''VI Olomoucké sympozium ukrajinistů střední a východní Evropy''. Olomouc, Česko, 21-23.08.2014. Olomouc: Univerzita Palackeho v Olomouci, pp.424-427.

Sokolov, A. and Savchenko, A.V., 2019. Voice command recognition in intelligent systems using deep neural networks. In: *2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI).* Herlany, Slovakia, [online] 24-26 January 2019, pp.113-116. IEEE. Available at: <https://doi.org /10.1109/SAMI.2019.8782755> [Accessed 21 April 2022].

Subbotin, S.O., Oliinyk, A.O. and Oliinyk, O.O., 2009. *Neiteratyvni, evoliutsiini ta multyahentni metody syntezu nechitkolohichnykh i neiromerezhnykh modelei* [Non-iterative, evolutionary and multiagent methods of synthesis of fuzzy and neural network models]. Zaporizhzhia: ZNTU.

Swamy, S., and Ramakrishnan, K.V., 2013. An efficient speech recognition system. *Computer Science & Engineering: An International Journal (CSEIJ),* [online] 3 (4), pp.21-27. Available at: <https://doi.org/10.5121/cseij.2013.3403> [Accessed 21 April 2022].

Systema rozpiznavannia holosu [Voice recognition system]. *Kia.* [online] Available at: <http://webmanual.kia.com/STD_GEN5W_8/AVNT/EU/Ukrainian/voicerecognitionsystem.html> [Accessed 15 April 2022].

Toda, T., Nakagiri, M. and Shikano, K., 2012. Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing,* [online] 20 (9), pp.2505-2517. Available at: <https://doi.org/10.1109/TASL.2012.2205241> [Accessed 22 April 2022].

Zheng, Y., Meng, Y. and Jin, Y., 2011. Object Recognition using Neural Networks with Bottom-up and Top-down Pathways. *Neurocomputing*, 74, pp.3158-3169.

Internet rechei [Internet of speeches]. [online] Available at: <https://iotukraine.com> [Accessed 25 April 2022].

UDC 004.032.26:004.8

Tkachenko Kostiantyn,
PhD in Economics,
Associate Professor at the Department of Information Technologies and Design,
State University of Infrastructure and Technology,
Kyiv, Ukraine
tkachenko.kostyantyn@gmail.com
https://orcid.org/0000-0003-0549-3396

Brusientsev Vladyslav,
Master's Student at the Department of Information Technologies and Design,
State University of Infrastructure and Technology,
Kyiv, Ukraine
vladbrusentcev1@gmail.com
https://orcid.org/0000-0002-8106-5855

USE OF NEURAL NETWORKS IN VOICE COMMANDS RECOGNITION

The purpose of the article is to research, analyze and consider general problems and prospects for the development of voice command recognition systems using the capabilities of neural networks, using the latest neural network technologies.

The research methodology consists in methods of semantic analysis of this subject area's basic concepts (voice command recognition systems). The existing systems and recognition algorithms are considered in the article.

The scientific novelty of the research is the analysis of modern voice recognition systems, the results of which can be used in the development of their own recognition system based on the use of improved speech models and recurrent neural network learners.

Conclusions. The efficiency of using neural networks for voice command recognition tasks is proved. Based on the research, a speech recognition system based on neural networks has been developed using an improved speech model.

Keywords: neural networks; neural network training; recurrent neural networks; discernment; artificial Intelligence; voice commands.

29.04.2022