УДК 004.032.26:[81'322.4+7.071.3
DOI: 10.31673/2412-9070.2024.022326
О. В. САЧУК, магістр;
В. А. САГАЙДАК, аспірант,
Державний університет інформаційно-комунікаційних технологій, Київ

РОЗРОБЛЕННЯ МЕТОДИКИ ТРАНСКРИБАЦІЇ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

У сучасному світі, де технології стрімко розвиваються, все частіше використовуються аудіо- та відеоматеріали для різноманітних цілей: навчання, досліджень, розроблень, а також у повсякденному спілкуванні та передаванні інформації. Цей підхід є зручним, ефективним та не відвертає нашої уваги від інших справ, на відміну від, наприклад, ручного письма або набору тексту на комп'ютері, який може бути незручним у певних ситуаціях, як-от під час поїздки в автомобілі. Водночас часто постає потреба перевести аудіо- та відеоінформацію в текстовий формат для її подальшого редагування, перенабору чи розшифрування, оскільки не всі здатні або мають змогу сприймати інформацію як звук чи відео. Окрім того, не кожен аудіо- чи відеозапис має чітку та зрозумілу вимову, що також зумовлює потребу в транскрибації — перекладі інформації з аудіо та відео в текстову форму. Одним із ключових аспектів цього процесу є розпізнавання важливої інформації, що є головним компонентом сучасного використання комп'ютерних технологій. Стосовно реалізації таких завдань людство постійно шукає нові підходи, серед яких особливе місце посідають нейронні мережі — складні алгоритмічні конструкції, здатні навчатися та адаптуватися до різноманітних типів даних для їх ефективного оброблення.

Ключові слова: транскрибація; нейронні мережі; методика транскрибації.

Вступ

Постановка проблеми. Нейронну мережу (НМ) можна подати як інформаційно-математичну програму, збудовану як біологічну нейронну мережу за прикладом мережі нервових клітин — мозку. Якщо за принцип дії брати аналогію мозку живих істот, то ключовим моментом буде виступати штучний нейрон як модель, що імітує нервову клітину мозку — біологічний нейрон. Термін «штучний нейрон» виник під час дослідження утворюваних у мозку процесів, зокрема в разі спроби поновити ці процеси за допомогою техніки. Першою такою спробою були нейронні мережі Маккалока і Піттса, якими зараз послуговуються в багатьох задачах прогнозування, для розпізнавання образів, у задачах керування тощо [1]. Це стало можливим після того, як науковці набули вміння навчати нейронні мережі розпізнавати картинки, текст, звуки.

Сучасна транскрибація перейшла від ручного методу, коли транскрибацію виконували люди, до майже повністю технічного методу, коли її здійснюють розумні машини, а саме нейронні мережі, тоді як людина лише коригує результат і вдосконалює його. Коригування також потрібно і в разі, якщо такі проблеми, як шум, нечітка дикція, незнайомі нейронній мережі слова, неправильний вибір мови стає перепоною до правильного розшифрування мовлення нейронною мережею. Оброблення усного мовлення, тобто сама транскрибація, полягає в процесі вивчення звукових сигналів, оброблені цих сигналів, а далі перероблені їх у текстових формат.

Розпізнавання мовлення полягає в тому, щоб подавати кожне слово окремою позначкою й узгоджувати з доступними позначками, які нейронна мережа бере із статистики та інтернету. Далі НМ визначає слово і може розпочинати його друкування. На противагу старомодному методу прихованої марковської моделі, НМ не потребують попередніх знань щодо мовленнєвого процесу, а також статистики мовних даних [1]. Що ж впливає на вимову диктора? Характеристики голосу, а саме: тональність, його висота, грубість, гучність, акцент; характеристики дикції: правильна вимова, пунктуація, паузи, чіткість вимови; фонові чинники: шум, відлуння, мікрофон і його характеристики. Також має вплив уважність диктора та просто людський чинник, ступінь знання диктора, про що він диктує, його здатність узагалі диктувати, його втома тощо.

Розроблення методики транскрибації на основі нейронних мереж буде актуальною, поки НМ технічно та правильно не стануть досконало виконувати транскрибацію без потреби коригування з боку людини.

Аналіз останніх досліджень та публікацій. Що є звуком взагалі? Звук — це хвилеподібний рух частинок у різних середовищах, що поширюється в рідини, твердих тілах чи газах та сприймається слуховим апаратом (мовлення, музика) з частотою від 20 до 20000 Гц. Також під звуком розуміють коливання, котрі відчуває сенсорно-слухова система людей та тварин (шум, гул, вібрація) [2]. Звук найчастіше виникає в процесі руху чогось: голосових зв'язок під час мовлення, вібрації струн

або повітря в духових інструментах під час музики, у разі вибуху, падіння, руйнування, стуку, під час шуму та гулу.

Людина сприймає відносно невеликий діапазон частот, є приклади, коли людині ставало зле і/або лячно через дуже високий/низький діапазон звуку (що іноді призводило до виникнення містичних історій). На відміну від людей тварини сприймають більш широкий діапазон звуку.

Звук також має властивість поширюватися, проте водночас він слабшає, а в разі будь-яких перепон (стіни, дерева, гори) може й зовсім зникнути, хоча спочатку і був дуже гучним. Таке слабшання звуку відбувається через силу пружності, що зупиняє вібрацію. Звук існує навіть тоді, коли його ніхто не сприймає, якщо вібрація відбулася, вона буде поширюватися і слабшати, поки не зійде нанівець. Якщо ніхто не чує звук, то він або зовсім зникне, або просто не вистачає можливості його чути, але він все одно є.

Вивчення закономірностей сприйняття, генерації та поширення звукових коливань у різноманітних середовищах належить до галузі наукових знань під назвою акустика. З погляду акустики шум, або акустичний шум — це нестійкі або випадкові акустичні коливання частинок довкілля, що сприймаються органами слуху людини як небажані сигнали, і характеризуються випадковою зміною амплітуди і частоти [3].

Майже всі природні явища супроводжуються певними звуками, що відчуваються та розпізнаються слуховими апаратами тварин та людей і слугують для спілкування та орієнтування у просторі [4]. Людина чує звуки і поділяє їх на гармонійні для себе та на дратівливі чи небажані, і особливістю сприйняття таких звуків є те, що для кожної людини це індивідуально.

Метою статті є розроблення методики транскрибації на основі нейронних мереж із подальшим дослідженням сучасних методів транскрибації.

Основна частина

За способом творення звуки мовлення поділяються нейронними мережами на два процеси:

- на основі процесу обструкції (звуків з урахуванням наявності звуків перепон та без них). Ця класифікація базується на уявленні про рух звуку в повітрі, зокрема на класифікації, чи це голосний звук (тобто утворюється без перепон, коли струмінь повітря, проходячи через голосові зв'язки, потрапляє в ротову порожнину і просто виходить голосним звуком, такий звук ще може проспівуватися), чи це приголосний (звук, що утворився за допомогою перепони в роті, такий звук проспівати неможливо, але він здатен утворювати гул);
- на основі процесу творення голосних (коли голосові зв'язки вібрують) та глухих (коли голосові зв'язки не вібрують) звуків. За цією класифікацією всі голосні звуки дзвінкі, а приголосні можуть буть і дзвінкими, і глухими.

Методи розпізнавання мовлення, застосовні НМ, можна розділити на такі категорії:

- виокремлене слово — процес цього мовлення вимагає вимовляти слова окремо, з чіткою дикцією й обов'язковими паузами між словами. Таке мовлення нейронні мережі сприймають більш чітко, ніж неперервне мовлення, і припускаються менше помилок або взагалі їх не роблять;
- безперервні слова — мовлення, коли слова вимовляють чітко, але одне за одним, іноді прибігаючи до правил вимовляння пунктуації, це більш природний метод диктування, але все одно саме диктування. Нейронні мережі біль-менш добре транскрибують таке вимовляння, проте можуть неправильно визначати межі висловлювань, через що плутати слова, неправильно їх розшифровувати. А якщо й дикція буде поганою, то слова плутатимуться, неправильно розшифровуватимуться або взагалі не будуть розпізнаватися, а отже, виводитися в тексті як вигуки;
- сполучені слова — диктування здійснюється висловлюваннями, але самі висловлювання промовляються природно;
- спонтанна мова — це навіть не диктування, а просто мовлення лекції і/або книги. Це мовлення нейронні мережі дуже рідко можуть повністю транскрибатизувати, оскільки лекція промовляється природно. Найчастіше заважає ще й дикція, незвичайні висловлювання, ігнорування правил пунктуації. Навіть вислови іншими мовами призведуть до неправильного розшифрування мовного тексту.

Саме поняття нейронної мережі та штучного інтелекту спрямовує на створювання механічного розпізнання, подібного до розпізнання інформації людиною. Людина використовує візуалізацію, аналізує здобуту інформацію і вирішує, що потрібно зробити з цією інформацією. Акустичну інформацію людина сприймає так само. Але для НМ потрібно спочатку систематизувати і промаркувати акустичну інформацію, щоб потім швидко її знайти. Для цього їй потрібні знання лексики, фонетики, семантики та синтаксису. Те, що людина сприймає як щось дане, НМ має систематизовувати і маркувати. Нейронна мережа вивчає взаємозв'язки між фонетичними подіями, а отже, репрезентує знання та інтегрує джерела знань [5].

Нейронні мережі вчені називають нейронами або ланцюгами. Зараз НМ називають штучною нейронною мережею, що становить штучні нейрони або вузли. Штучні нейрони діють як і біологічні — отримують на вході інформацію, змінюють стан відповідно до неї та генерують вихідну інформацію. НМ робить це на основі математичної моделі

для оброблення інформації, що використовує конекціоністський підхід до спілкування. Нейронні мережі є простими математичними моделями, які визначають функцію $f: X \to Y$, або розподіл за X, або обидва X і Y, які постійно навчаються.

Транскрибація тільки засобами НМ може давати повністю автоматизований засіб транскрибації, але він має обробляти широкий спектр мовленнєвих функцій.

Можна виокремити такі методи транскрибації:

• ***фонетична*** — процес перетворення звуків мови на символи для їх точнішого подання в текстовому вигляді. Застосовується, наприклад, у процесі навчання голосових помічників: Google Assistant, Siri від Apple. Допомагає розпізнавати та інтерпретувати різні вирази та акценти, щоб дати більш точні відповіді на запити користувача. Використовується також під час аналізу музики: щоб ідентифікувати артикуляцію музичних звуків та виразні особливості твору;

• ***графематична*** — запис слів, як вони пишуться літерами. Використовується в автоматичних перекладачах, системах розпізнавання мови, застосунках для людей із порушеннями слуху з перетворенням мови в текст тощо. Можливі проблеми з контекстом: багато слів пишуться однаково, але мають різне значення;

• ***лінгвістична*** — заснована на лінгвістичному аналізі мови та її звуковій структурі. Цей метод транскрибації може бути корисним для аналізу акценту та вивчення мовних лінгвістичних особливостей. Зокрема, щоб досліджувати, як змінюються звуки залежно від особливостей вимови носія.

Загалом, графематична транскрибація перетворює усне мовлення на текст буквально, не беручи до уваги вимову. Фонетична зважає на всі деталі вимови, включно з фонетичними особливостями. Наприклад, слово «bath» в українській транскрибації можна записати графематично як «бат», але фонетично воно матиме вигляд як [bæθ]. А отже, щоб врахувати різні діалекти і допомогти людині, яка не знає англійської мови, потрібно правильно вимовити слово.

Зі свого боку, лінгвістична транскрибація також використовується для запису вимови слова, але може бути абстрактнішою, сконцентрованою на фонологічних відмінностях мови, а не на конкретних звуках.

Залежно від цілей користувача для забезпечення транскрибації застосовуються різні нейронні мережі і/або різні програми.

Варіантів багато, один із них — Google Cloud Speech-to-Text API, транскрибує мову в текст за допомогою глибоких нейронних мереж. Забезпечує високу точність розпізнавання, підтримує багато мов та форматів аудіофайлів. API можна інтегрувати в мобільні програми, розумні пристрої, системи спостереження тощо. Наприклад, у телефонному автовідповідачі Cloud Speech-to-Text оброблятиме голосові повідомлення від клієнтів і перетворюватиме їх на текст. У лікарняному устаткуванні технологія здатна допомогти медперсоналу записувати діагнози та схеми лікування до карти клієнта голосом.

Ще один сервіс — Amazon Transcribe, який використовує методи глибокого навчання, рекурентні нейронні мережі (RNNs). Виконує не тільки транскрибацію, але також розпізнає та розділяє уривки за спікерами. Наприклад, ви не змогли бути присутніми на важливій конференції і тому придбали аудіозапис. Але без тексту ви витратите багато часу на прослуховування. Amazon Transcribe допоможе вирішити цю проблему: у сервіс можна завантажити аудіо, і він автоматично перетворить мову в текстовий формат. Розшифровку можна швидко перегорнути, знайти потрібні уривки за ключовими словами (якщо ми говоримо про електронну версію) і зробити нотатки, не гаючи часу на прослуховування всього запису.

Бібліотека Kaldi з відкритим вихідним кодом використовує методи глибокого навчання та статистичних моделей для транскрибації аудіофайлів. Вона застосовна, зокрема, в Amazon Alexa та Google Assistant для перетворення голосових команд у текст, у програмному забезпеченні для розшифрування мовлення в реальному часі на телебаченні.

Висновки

• Транскрибація підвищує доступність вмісту. Так, наприклад, відео з субтитрами можуть дивитися люди з порушенням слуху. Також субтитри допомагають дивитися фільми в оригіналі, зокрема в галасливій обстановці.

• Текстові стенограми аудіоконтенту (подкастів, радіопрограм) допомагають тим, хто не має можливості прослухати інформацію тієї ж миті, але хотів би з нею ознайомитися.

• Транскрибація слугує допоміжним інструментом під час навчання. Наприклад, штучний інтелект швидко перетворить у текст записи лекцій та семінарів, хоч іноді і потребуватиме після цього коригування. На друкованій копії студент може робити позначки, залишати коментарі.

• Транскрибації ділових зустрічей, дзвінків чи конференцій можна роздавати співробітникам для обговорення, передавати відсутнім.

• Використання транскрибації в медицині та охороні здоров'я дає змогу розшифровувати розмови з пацієнтом, швидко знаходячи потрібне місце в розмові, і ще раз проаналізувати симптоми. Текстовий запис сеансу терапії у психолога використовується, наприклад, щоб відстежити прогрес.

• Транскрибація покращує видимість контенту в пошукових системах. Якщо додати до відео та аудіо транскрибацію, матеріал простіше знайти за ключовими словами, його позиції в пошуковій видачі піднімаються.

• Транскрибація також здатна легко перепрофільовувати контент, зокрема, перероблювати ролик на пост у блозі, удосконалюючи «розумні» пристрої.

• Автоматичне розпізнавання мови використовується у величезній кількості приладів, від Телеграма та Вайбера до медичних пристроїв для людей з обмеженими можливостями, наприклад, автоматизовані системи для керування інвалідним кріслом за допомогою голосових команд, комунікатори для людей з афазією. Технологія застосовується і в робототехніці для створення роботів, які рухаються, думають та говорять як люди.

Методика транскрибації залежно від цілей користувача може охоплювати користування різними сервісами і комбінування варіацій транскрибації.

Список використаної літератури

*1. **Zou J., Han Y., So S. S.** Overview of Artificial Neural Networks. In: Livingstone D.J. (Eds) Artificial Neural Networks // Methods in Molecular Biology™. 2008. Vol. 458. P. 14–22.*

*2. **Голямина И. П.** Звук. Физическая энциклопедия. URL: www.femto.com.ua/articles/part_1/1222.html*

*3. **Шум.** Вікіпедія. URL: https://uk.wikipedia.org/wiki/Шум*

*4. **GetAClass** — Физика в опытах и экспериментах. Источники звука. URL: https://youtu.be/nFcmTTT9yiE*

*5. **Suryo Wijoyo.** Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot // International Conference on Information and Electronics Engineering IPCSIT. Singapore, 2011. Vol. 6, IACSIT Press. P. 179–183.*

O. Sachuk, V. Sahaidak

DEVELOPMENT OF A NEURAL NETWORK-BASED TRANSCRIPTION METHOD

In today's ever-evolving digital era, the utilization of audio and video materials has become increasingly prevalent in various spheres, including educational sectors, research initiatives, technological developments, and daily communication. This trend is driven by the convenience and efficiency of these mediums, offering quick access to information without the distractions often associated with more traditional methods like manual writing or typing, especially in scenarios where such activities are impractical, such as when traveling in a vehicle.

However, this shift towards audio-visual content has introduced the challenge of converting these dynamic formats into text for a range of purposes. This need arises for several reasons. Firstly, editing and refining spoken content is often easier when it's transformed into a written format, allowing for a more thorough review and adjustment process. Secondly, in situations where individuals prefer or require written documentation — for instance, in academic or professional settings — transcription becomes a critical tool. Additionally, the clarity of audio and video content can vary greatly, with factors like diction, accent, background noise, and recording quality affecting the comprehensibility of the material.

To address these challenges, transcription — the process of translating audio and video content into text — has become a valuable solution. It involves a meticulous process of listening, interpreting, and typing out the content, ensuring that the essence and nuances of the original material are accurately captured. This task demands not only attention to detail but also a deep understanding of the context and subject matter to ensure precision and reliability.

In the realm of technology, advancements such as neural networks have revolutionized the transcription process. Neural networks, complex algorithmic structures that emulate human brain functioning, can learn, adapt, and process diverse types of data. In transcription, they are employed to recognize and interpret various speech patterns, accents, and languages, significantly enhancing the accuracy and speed of converting spoken words into written text. This integration of advanced technology in transcription not only streamlines the process but also opens new possibilities for accessibility, research, and data analysis, making it an indispensable tool in the modern world.

Keywords: transcription; neural networks; transcription methodology.