УДК 004.93

# ПІДХІД ДО АВТОМАТИЗАЦІЇ КОНВЕРСІЇ ВІДЕО ТА АУДІО МАТЕРІАЛІВ У ТЕКСТ

Шимко Д.І., Левикін І.В.
e-mail: dmytro.shymko@nure.ua, ihor.levykin@nure.ua
Харківський національний університет радіоелектроніки, каф. МСТ
м. Харків, Україна

The automation of media content conversion, particularly video-to-text transformation, presents challenges related to accuracy and information loss. Automated transcription systems often struggle with factors such as audio quality, speech clarity, and domain-specific terminology. A two-step approach, combining initial rough transcription with AI-driven contextual enhancement, improves recognition accuracy. The first stage generates a draft transcript, while the second refines it using additional metadata and linguistic models. Contextual adaptation significantly enhances precision. Although this method increases processing costs, it optimizes transcription quality, making automated media conversion more reliable and applicable across various domains.

Автоматизація процесу конверсії медіа-матеріалів може включати велику кількість етапів, кожен з яких впливає на кінцеву якість отриманого тексту. Деякі типи перетворення, зокрема трансформація відео у текстовий формат, можуть призводити до втрати первинних характеристик контенту, оскільки текст не здатний повністю передати графічні та аудіальні особливості відео. Додатково, у процесах, таких як транскрипція аудіо, досягти ідеальної точності майже неможливо, якщо матеріали не створені в ідеальних умовах (студійний запис, шумоізоляція, чітка вимова, використання стандартної лексики, англійська мова тощо).

Перетворення відео у текстовий формат без попередньо підготовлених субтитрів характеризується обмеженою якістю, що може спричинити значну кількість артефактів та помилок. Це зумовлено такими факторами, як якість звуку, чіткість вимови, якість запису голосу та наявність фонових шумів, наприклад музики або сторонніх звуків. Окрім того, швидкість мовлення, акцент та використання діалектів також можуть впливати на якість транскрипції, ускладнюючи процес автоматичного розпізнавання.

Точність розпізнавання тексту значною мірою залежить від наявності технічних або професійних термінів. Автоматизовані системи, що ґрунтуються на нейромережах, демонструють результати, які залежать від якості навчальних даних. Рідковживані слова, зокрема технічні терміни, географічні назви та абревіатури, часто мають низьку точність розпізнавання через недостатнє представлення у навчальних вибірках. Це особливо актуально для спеціалізованих галузей, таких як медицина, інженерія або наукові дослідження, де важливими є точність і правильна інтерпретація термінології.

Деякі сучасні сервіси та моделі штучного інтелекту дозволяють передавати не лише аудіо- або відеофайл, а й додатковий контекст, який може містити інформацію про тематику відео, використані терміни та ключові слова. У випадку ручного транскрибування оператор має можливість прослухати аудіозапис і створити детальний опис контексту, що істотно підвищує якість перетворення. Однак в умовах автоматизації виникають труднощі у формуванні контексту, оскільки неможливо створити універсальне правило, яке б однаково ефективно працювало для різних типів контенту. Тому адаптація алгоритмів розпізнавання під конкретні задачі є критично важливою для підвищення точності результатів.

Метою дослідження є підвищення ефективності та якості транскрипції відео та аудіо матеріалів у випадку автоматизації цього процесу, щоб уникнути залучення оператора для мануального контролю та створення додаткових даних про матеріали. Актуальність дослідження зумовлена необхідністю уникнення помилок при конверсії медіа матеріалів, що може призвести до подальших проблем з обробкою та використанням отриманих медіа матеріалів, які можуть мати у собі спотворену та неправдиву інформацію та дані [1].

Для вирішення зазначених проблем доцільно використовувати такий підхід до обробки аудіо, за якого обробка матеріалу повторюється після отримання першого результату та його додаткової обробки, після чого покращені дані знову відправляються на опрацювання.

На першому етапі створюється чорновий варіант тексту, який може містити помилки та неточності. Незважаючи на це, застосування генеративних моделей штучного інтелекту дозволяє виявити основні особливості тексту, включаючи ключові слова, терміни, імена, назви компаній та географічні об’єкти. Це дає змогу формувати контекст, який на другому етапі сприяє точнішому розпізнаванню складних слів та спеціалізованих термінів. Додатково можлива обробка тексту на рівні граматики та стилістики, що підвищує його читабельність та відповідність оригінальному змісту.

На другому етапі створений контекст разом із файлом аудіозапису передається до сервісів транскрипції, що значно підвищує якість розпізнавання за рахунок точнішого визначення тематики відео та використаних термінів. Додатково до контексту можна передавати метаінформацію, таку як назва та опис відео або аудіозапису (наприклад, у випадку відео, опублікованого на платформах на кшталт YouTube), а також інші дані, надані авторами контенту. Крім того, використання алгоритмів післяобробки тексту дозволяє усунути неточності, пов'язані з автоматичним розпізнаванням мовлення, наприклад граматичні помилки або неправильні розділові знаки.

Запропонований підхід суттєво покращує якість конверсії медіафайлів у текстовий формат, хоча й підвищує витрати на обробку через необхідність додаткових запитів до сервісів, що використовують генеративні моделі штучного інтелекту. Враховуючи це, доцільно оцінювати фінансовий фактор такого підходу. Оптимізація витрат можлива шляхом використання менш затратних сервісів для первинної обробки [2], оскільки навіть частково сформований контекст на другому етапі сприяє значному підвищенню якості транскрипції.

В результаті проведеного дослідження було здійснено конверсію відео та аудіо матеріалів у текст шляхом транскрипції з використанням сервісів із моделями генеративного штучного інтелекту та проведено порівняння кількості помилок і неточностей у випадку звичайної одноразової транскрипції без передачі додаткового контексту. Було визначено, що передача додаткового контексту про сутність контенту в процесі створення транскрипції дозволяє зменшити кількість помилок і неточностей порівняно зі звичайною транскрипцією. Такий підхід є найбільш ефективним у випадках обробки матеріалів, у яких використовуються рідковживані слова, такі як визначення, терміни, абревіатури або географічні назви. Підвищення ефективності досягається шляхом додаткового уточнення під час передачі контексту про сферу діяльності, можливі терміни, назви, об’єкти та інші деталі, що дозволяє моделям генеративного штучного інтелекту краще розуміти зміст матеріалу та використані слова.

Подальший розвиток алгоритмів розпізнавання мовлення та їх адаптація до специфічних задач не лише підвищить точність автоматичних транскрипцій, але й зробить процес конверсії медіа-матеріалів більш ефективним для конкретних сфер, таких як журналістика, освіта, наукові дослідження, медіа-аналітика та створення субтитрів, забезпечуючи мінімізацію втрат змісту та коректну інтерпретацію спеціалізованої термінології.

Список використаних джерел:

1. Шимко Д.І., Штанько В.І. (2025). Виклики збереження сутності та автентичності при автоматизації процесу конверсії медіа-матеріалів // Міжнародна науково-практична конференція молодих учених, аспірантів та студентів "Інформаційні технології в сучасному світі: дослідження молодих вчених": тези доповідей, 27 – 28 лютого 2025 р. Х.: ХНЕУ імені Семена Кузнеця, (с. 207).

2. Levykin, V., Ievlanov, M., Neumyvakina, O., Levykin, I., & Nakonechnyi, A. (2024). Estimation of IT-project efforts for information system creation in the conditions of re-use of its functions. Information technology. Industry control systems. T. 2, 2(128), 6-19. https://doi.org/10.15587/1729-4061.2024.301227.