УДК 004.934:81'322.4

АНАЛІЗ НЕЙРОМЕРЕЖЕВОГО КОНВЕЄРА ПЕРЕКЛАДУ МОВЛЕННЯ ДЛЯ ОНЛАЙН-НАВЧАННЯ

Довгун Д.С.
Національний університет «Одеська юридична академія»
Україна, Одеса
dovgundovgundovgun@gmail.com

Анотація. *Проаналізовано реалізацію нейромережевого конвеєра перекладу усного мовлення в режимі, близькому до реального часу, інтегрованого у вебплатформу для онлайн-навчання. Конвеєр охоплює розпізнавання мовлення (OpenAI Whisper), машинний переклад та синтез мовлення (edge-tts). Затримка та якість перекладу визначаються переважно етапом розпізнавання, зокрема розміром моделі та обчислювальними ресурсами сервера, а етап машинного перекладу знижує точність передавання контекстно залежної лексики. Сформульовано практичні рекомендації щодо вибору моделей і апаратних ресурсів для впровадження подібних систем в освітніх платформах.*
Ключові слова: *машинний переклад; розпізнавання мовлення; синтез мовлення; нейронні мережі; онлайн-навчання.*

Вступна частина. За даними [1], приблизно 49% студентів у світі певною мірою залучені до онлайн-навчання. Пандемія COVID-19 прискорила поширення цього формату, а платформи Google Meet, Zoom та Microsoft Teams стали альтернативою очним заняттям [2]. Онлайн-формат уможливлює проведення занять незалежно від відстані, безпекової ситуації та фізичних обмежень. Це відкриває доступ до міжнародних курсів, конференцій і спільних наукових проєктів. Водночас мовний бар'єр залишається одним із головних чинників, які обмежують участь у такій взаємодії, особливо коли йдеться про професійну термінологію. Послуги синхронних перекладачів переважно платні й не завжди доступні, тому автоматичний переклад мовлення, який знижує цей бар'єр без їхньої участі, є практично значущим для освітніх закладів.

Розвиток штучного інтелекту та машинного навчання сприяв появі інструментів автоматичного перекладу мовлення. Режими Conversation і Live translate в Google Translate [3] перекладають мовлення співрозмовників по черзі. Розширення для браузера JotMe [4] пропонує переклад і субтитри для онлайн-конференцій у режимі, близькому до реального часу, а LiveLingo [5] орієнтований на телефонні дзвінки. Основними недоліками таких інструментів є обмежена здатність враховувати контекст, затримки та чутливість до якості аудіосигналу. Крім того, це переважно закриті комерційні рішення, тому їхня внутрішня архітектура недоступна для аналізу й не може бути безпосередньо адаптована до потреб освітньої платформи.

Проблематика полягає в тому, що для освітніх сценаріїв бракує систематизованих даних про те, який етап конвеєра «розпізнавання → переклад → синтез» найбільше впливає на затримку та на якість результату і які апаратні ресурси потрібні для прийнятної роботи. Без цього неможливо обґрунтовано вибрати моделі та конфігурацію сервера, а помилки на початковому етапі каскадно поширюються на наступні.

Метою роботи є аналіз реалізації перекладу мовлення в режимі, близькому до реального часу, на вебплатформі для навчання, виявлення причин недоліків, що проявилися під час тестування, та обґрунтування шляхів їх усунення.

Основна частина. Програмний застосунок розроблено автором на базі вебплатформи Ew/oBs (Education without Borders) з використанням мікросервісної архітектури. Систему

поділено на два сервери, реалізовані за допомогою фреймворка FastAPI: головний сервер і сервер ШІ-моделей. Головний сервер виконував маршрутизацію сторінок, керування навчальними класами, оброблення завантажених файлів і машинний переклад тексту. Сервер ШІ-моделей відповідав за переклад усного мовлення за допомогою конвеєра нейромережевих моделей. Тестування проводилося для напрямів [UA→EN, EN→UA] на 20 фрагментах аудіозаписів тривалістю від 5 до 30 секунд у конфігурації AMD Ryzen 5 3600, NVIDIA GeForce RTX 2060, Kingston FURY DDR4-3200.

Першим етапом було розпізнавання мовлення. Запис аудіо ініціювався на стороні клієнта функцією Push-To-Talk, після чого аудіозапис разом із цільовою мовою перекладу надсилався на сервер ШІ-моделей. Модель OpenAI Whisper [6] виконувала транскрибування й повертала текст. Під час тестування саме цей етап спричиняв більшість проблем. Whisper має декілька розмірів моделей: tiny, base, small, medium, large-v3 та large-v3-turbo, і від вибраного розміру суттєво залежать швидкість і якість транскрибування. Модель large-v3-turbo обробляє 30-секундний фрагмент приблизно за 0,5–1 с, зберігає контекст, коректно транскрибує складну термінологію та є стійкою до низької якості звуку, проте потребує GPU рівня не нижче NVIDIA RTX 3050. Модель small можна запускати на слабких процесорах, і вона швидко обробляє аудіо, але допускає значну кількість помилок, може генерувати хибний текст на ділянках тиші та чутлива до акустичних умов. Найвищу якість забезпечує large-v3-turbo, що зменшує кількість помилок на наступних етапах, однак для її стабільної роботи потрібен сервер із GPU рівня NVIDIA RTX 3050 або вище.

Другим етапом був машинний переклад, реалізований за допомогою бібліотеки deep-translator з підключенням до Google Translate. Текст, отриманий після транскрибування, передавався до модуля машинного перекладу. Формальну мову система перекладала якісно, проте контекстно залежну лексику та спеціальну термінологію часто передавала некоректно, а розмовні й сленгові вирази перекладала дослівно. Помилки першого етапу також безпосередньо впливали на результат: на рис. 1 хибно розпізнану власну назву (Godalming, Surrey → «Godelming, Syria») було передано у перекладі як є.

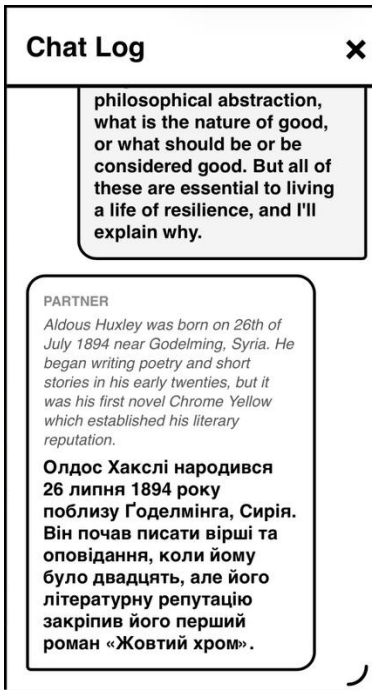


Рис. 1. Приклад результату роботи конвеєра: розпізнаний текст (Whisper medium) і його переклад (Google Translate)

Третім етапом був синтез мовлення, реалізований за допомогою бібліотеки edge-tts [7], що використовує нейромережеві голоси від Microsoft. Перекладений текст озвучувався нейронним голосом із природною інтонацією та паузами. Аудіо передавалося до браузера співрозмовника через WebSocket і відтворювалося, а переклад дублювався текстом у чаті (рис. 1).

Висновки. Переклад мовлення в режимі, близькому до реального часу, складається з трьох етапів: транскрибування, машинного перекладу та синтезу мовлення, і для кожного з них визначено відповідні нейромережеві моделі. Головні недоліки, а саме затримка, низька якість перекладу та залежність від умов аудіо, виникають переважно на етапі розпізнавання через недостатню обчислювальну потужність сервера й неоптимальний вибір розміру моделі. Додатковим джерелом помилок є етап машинного перекладу, який недостатньо враховує контекст. Практична цінність роботи полягає в рекомендаціях щодо вибору моделі та апаратних ресурсів для освітніх платформ: для стабільної якості доцільно використовувати модель large-v3-turbo на сервері з GPU рівня RTX 3050 або вище, а моделі малого розміру придатні лише для обмежених ресурсів з очікуваним зниженням точності.

Використання інструментів ШІ. Claude (Anthropic) використано для перевірки тексту на помилки та для редагування формулювань.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Kathakali B. (2025). Online learning statistics: is online education the future? *Atlantic International University*. https://www.aiu.edu/blog/online-learning-statistics-future-education/

[2] Shabir, A., Kamran, S. (2022). Comparative Study of Alternative Teaching and Learning Tools: Google Meet, Microsoft Teams, and Zoom during COVID-19. In: *Teaching in the Pandemic Era in Saudi Arabia*, 120–129. https://doi.org/10.1163/9789004521674_008

[3] Sheets, M. (2025, August 26). *New Ai-powered live translation and language learning tools in Google Translate.* News from Google. https://blog.google/products-and-platforms/products/translate/language-learning-live-translate/

[4] JotMe. (2026). *One platform across languages*. https://www.jotme.io/

[5] LiveLingo. (2026). *Watch anything. Understand everything*. https://getlivelingo.com/

[6] OpenAI. (2022). *Introducing Whisper*. https://openai.com/index/whisper/

[7] Edge TTS. (2026). *Free text to speech online - 74 languages, 322 voices*. https://edge-tts.com/

Dovhun D.S.

Analysis of a neural pipeline for speech translation in online education

Abstract. *This paper analyses a neural pipeline for near-real-time speech translation integrated into an educational web platform for multilingual online conferences. The pipeline combines speech recognition (OpenAI Whisper), machine translation and speech synthesis (edge-tts). The results show that latency and translation quality are determined mainly by the recognition stage, namely model size and server hardware, while machine translation reduces accuracy for context-dependent vocabulary. Practical recommendations are given on selecting models and hardware for deploying such systems in educational platforms.*

Keywords: *machine translation; speech recognition; speech synthesis; neural networks; online learning.*