УДК 004.8
DOI: https://doi.org/10.17721/3041-2323.2025.137-164

Владислав ЛУЦ, асп.
ORCID ID: 0009-0001-2948-6935
e-mail: tibet.septim@gmail.com
Національний транспортний університет,
Київ, Україна

Олександр БЕЗВЕРХИЙ, д.ф.-м.н., професор
ORCID ID: 0000-0002-0834-6335
e-mail: o_bezver@ukr.net
Національний транспортний університет,
Київ, Україна

Євгеній ТОПОЛЬСЬКОВ, к.т.н., доц.
ORCID ID: 0000-0001-5587-3069
e-mail: y.topolskov@knu.ua
Київський національний університет
імені Тараса Шевченка, Київ, Україна

ПРОГРАМНА СИСТЕМА ЗІ ШТУЧНИМ ІНТЕЛЕКТОМ ДЛЯ ПЕРЕКЛАДУ УКРАЇНСЬКОЇ ВЕРБАЛЬНОЇ МОВИ НА ЖЕСТОВУ МОВУ

У роботі подано програмну систему, призначену для перекладу української мови на анімації жестової мови у режимі реального часу. Система використовує модель Whisper для автоматичного розпізнавання мови, забезпечуючи надійну транскрипцію усної української мови. Оскільки українська мова є високофлексійною, а жестова мова не використовує граматичні відмінки, реалізовано модуль виправлення на основі JSON. Він охоплює два компоненти: словник для частих помилок розпізнавання та механізм лематизації, який переписує слова у їх базову форму. Нормалізований текст потім відображається на GIF-анімаціях, кожна з яких представляє відповідний жест української жестової мови. Процес перекладу відбувається в реальному часі, забезпечуючи послідовне відображення анімацій, синхронізованих із розпізнаною мовою. Рішення демонструє, як технічну ефективність (низьку затримку, модульну структуру), так і соціальний вплив через покращення доступності комунікацій для людей з вадами слуху чи тих, хто використовує жестову мову як основний засіб спілкування. Потенційні застосування охоплюють інклюзивну освіту, повсякденні

взаємодії та розроблення технологій, адаптованих до потреб спільноти з вадами слуху та мовлення.

Ключові слова: ***інклюзивні та асистивні технології, українська жестова мова, штучний інтелект, автоматичне розпізнавання мовлення, модель Whisper, лематизація, корекція помилок на основі JSON, GPU CUDA прискорення.***

Вступ

Технології інклюзивної комунікації стали ключовою сферою досліджень і розробок, особливо в контекстах, де усна мова недоступна для частини суспільства. В Україні тисячі осіб використовують українську жестову мову (УЖМ) (UTOG, 2020) як основний або бажаний спосіб комунікації. Однак системи перекладу реального часу, що заповнюють прогалину між усною українською та УЖМ, залишаються недостатньо розвиненими. Наявні інструменти транслювання мовлення у текст надають часткові рішення, але не враховують унікальну лінгвістичну структуру української мови чи специфічні вимоги представлення жестової мови.

Флексійні мови, такі як українська чи польська, створюють унікальні виклики для оброблення природної мови (NLP) через їх багаті морфологічні структури (Prystupa & Sydorenko, 2019). На відміну від аналітичних мов, зокрема, англійська, які значною мірою враховують порядок слів та допоміжні слова, флексійні мови використовують численні афікси, закінчення та внутрішні зміни слів для кодування граматичних категорій, таких як відмінок, рід, число, час і вид (Krak et al., 2017), що приводить до значно більшого набору можливих форм слів для однієї леми, створюючи труднощі для завдань, таких як токенізація, лематизація та машинний переклад.

Українська мова становить значний виклик через свою складну граматичну систему, яка охоплює численні відмінки, флексії та форми слів. Оскільки УЖМ не використовує граматичні відмінки, прямий переклад сирих транскриптів мови часто призводить до неправильних інтерпретацій або неприродних послідовностей жестів. Для усунення цієї прогалини доцільно здійснювати автоматичне розпізнавання

мови з інтелектуальною нормалізацією тексту, забезпечуючи перетворення слів у їх базові форми перед відображенням у форми жестів.

Ця робота має за мету продемонструвати технічну можливість поєднання передових систем розпізнавання мовлення з лінгвістичним обробленням для здійснення синхронного перекладу, а також висвітлити його практичне значення для підвищення рівня інклюзивності для осіб, які використовують жестову мову. Вона не лише сприяє повсякденній комунікації, а й відкриває нові можливості у сфері освіти, публічних послуг та асистивних технологій.

У дослідженні запропоновано програмне забезпечення з модульною структурою, що використовує модель штучного інтелекту Whisper для розпізнавання і точної транскрипції мовлення у поєднанні з інструментами корекції на основі формату JSON для усунення типових помилок розпізнавання та граматичних варіацій. Нормалізовані вихідні дані динамічно пов'язуються з бібліотекою GIF-анімацій, кожна з яких представляє окремий жест. Завдяки послідовному відтворенню цих анімацій система забезпечує зв'язний переклад українського мовлення жестовою мовою в режимі реального часу.

Огляд наявних досліджень

Дослідження і розробки автоматичного перекладу з усної мови на жестову мову постійно розвиваються протягом останніх двох десятиліть, але залишаються недостатньо глибоко опрацьованими порівняно з добре відомими технологіями транскрипції мовлення у текст або машинного перекладу між усними мовами. Запропоновано різноманітні підходи з різним акцентом на точність розпізнавання мови, моделювання граматики жестової мови, техніки рендерингу анімацій та оптимізацію продуктивності реального часу. Галузь зазнала значних технологічних зсувів – від систем на основі правил до статистичних підходів і нейронних архітектур, машинного навчання.

Ранні спроби перекладу мовлення на жести зосереджувалися переважно на текстових конвеєрах, де розпізнане мовлення спочатку транскрибувалося за допомогою конвенційних систем автоматичного розпізнавання мови, а потім переносилось на заздалегідь визначені глоси чи символічні представлення окремих жестів. Такі розроблені системи (Speers, 2001), (Zhao et al., 2000) встановили фундаментальні архітектури, але були обмежені обчислювальними ресурсами та примітивними можливостями оброблення природної мови. Також ці піонерські системи часто не мали складних механізмів лінгвістичної адаптації для високофлексійних мов, зокрема, українська чи інші слов'янські мови зі складними морфологічними системами. Крім того, багато ранніх проєктів мали суттєві обмеження через невеликий обсяг словникового запасу (зазвичай менше ніж 500 жестів), обчислювальні труднощі, що перешкоджали роботі в режимі реального часу, а також неналежне опрацювання фундаментальних граматичних відмінностей між усним і жестовим мовленням, що значно знижувало їхню придатність для використання в реальних комунікативних сценаріях.

Розвиток складніших технологій розпізнавання мови сприяв прогресу у галузі. Системи, подібні до розроблених (Stein et al., 2006) для німецької жестової мови та іспанської жестової мови (López-Ludeña et al., 2014), продемонстрували покращену точність розпізнавання завдяки кращому акустичному моделюванню та адаптаціям, специфічним для мови. Однак ці системи все ще стикалися з викликами обробки морфологічно багатих мов та досягнення продуктивності реального часу, необхідної для природного потоку комунікації.

Розроблення досконаліших технологій розпізнавання мовлення значно посприяло прогресу у цій галузі. Вище вказані системи продемонстрували підвищену точність розпізнавання завдяки покращеному акустичному моделюванню та адаптації до специфіки конкретних мов.

З появою архітектур машинного навчання та моделей на основі трансформерів інструменти розпізнавання мовлення зазнали удосконалень. Сучасні системи, такі як Vosk (Alphacephei), що базується на статистичній структурі Kaldi з нейромережевими покращеннями, та Whisper від OpenAI, яка використовує масштабні архітектури трансформерів, навчені на величезних багатомовних наборах даних, кардинально покращили можливості розпізнавання мовлення багатьма мовами, включаючи мови з обмеженими ресурсами.

Серед цих сучасних рішень Whisper вирізняється своєю стійкістю у складних акустичних умовах, багатомовними можливостями, що охоплюють понад 90 мов, і особливою адаптивністю до мов з обмеженими цифровими ресурсами, зокрема до української. Навчання моделі на різноманітних аудіоджерелах – від професійних записів до нефільтрованого онлайн-матеріалу – забезпечило позитивні результати в реальних умовах, де часто зустрічаються фоновий шум, кілька мовців та акустичні варіації. Хоча Whisper було всебічно вивчено та впроваджено для завдань транскрибування, машинного перекладу та сервісів доступності контенту, її специфічне застосування в контексті перекладу жестовою мовою – особливо для української жестової мови та інших жестових мов Східної Європи – залишається значною мірою недослідженим у наукових публікаціях.

Щодо представлення та візуалізації вихідних даних жестової мови, дослідницька спільнота переважно дотримувалася двох відмінних, але взаємодоповнюючих стратегій: систем анімації на основі аватарів і підходів з використанням попередньо записаних автентичних медіафайлів. Системи на основі аватарів, прикладом яких є проєкти EMBR (Elliott et al., 2008), eSign (Efthimiou et al., 2009) та роботи (Othman and Jemni, 2012), забезпечують значну гнучкість у створенні нових комбінацій жестів та опрацюванні раніше невідомої лексики за допомогою параметричних моделей анімації. Теоретично ці системи можуть генерувати будь-яку

послідовність жестів і адаптуватися до різних стилів виконання або регіональних варіантів. Однак вони мають кілька критичних обмежень: неприродні патерни руху, які не здатні вловити плавну динаміку людської жестової мови; труднощі у відтворенні тонких, але вирішальних аспектів, таких як міміка та немануальні маркери; а також відсутність культурної специфіки у представленні автентичних для спільноти стилів жестів. Нещодавні досягнення в нейронній анімації та синтезі руху почали вирішувати деякі з цих проблем, але системи на основі аватарів все ще намагаються досягти того рівня природності та культурної автентичності, якого потребують користувачі з вадами слуху для ефективної комунікації.

На противагу цьому, системи, що використовують попередньо записані зображення, відеопослідовності або анімовані GIF-файли, створені носіями жестової мови, досягають значно вищої автентичності та культурної відповідності результату. Проєкти, на основі SignWriting, розроблені (Sutton, 2019), та системи відеоперекладу, впроваджені (Dreuw et al., 2007), демонструють ефективність цього підходу у створенні лінгвістично точних і культурно адекватних представлень жестів. Проте ця автентичність досягається ціною необхідності створення великих лексичних баз даних, значних вимог до обсягу пам’яті для зберігання медіафайлів високої якості та зниженої гнучкості при роботі з новою лексикою або адаптації до різних регіональних варіантів жестів. Створення баз жестів потребує суттєвої співпраці зі спільнотами глухих, професійними лінгвістами та носіями мови, що робить розробку системи ресурсомісткою та тривалою у часі.

Мультимодальні нейронні архітектури – це перспективний напрям досліджень. Впровадження механізмів уваги (attention mechanisms) та архітектур трансформерів щодо перекладу жестової мови уможливило складніші підходи до обробки складних просторово-часових зв’язків, властивих жестовим мовам. Праці (Camgoz et al., 2020) та (Saunders et al., 2020) щодо перекладу жестової мови за принципом «послідовність до

послідовності» (sequence-to-sequence) та розробка підходів нейронного машинного перекладу, спеціально адаптованих для жестової мови, продемонстрували потенціал систем наскрізного навчання (end-to-end learning systems), здатних охоплювати як лінгвістичні, так і візуальні аспекти комунікації жестовою мовою. Ці підходи почали вирішувати деякі фундаментальні проблеми крос-модального перекладу, зберігаючи при цьому обчислювальну ефективність, придатну для застосунків реального часу.

Останніми роками увагу привернули дослідження, що стосуються безпосередньо морфологічно складних мов у контексті перекладу жестовою мовою. Хоча в галузі загальної обробки природної мови (NLP) є значні напрацювання з обчислювальної морфології для слов'янських мов, специфічні виклики адаптації морфологічно багатого усного введення до структур жестової мови стають об'єктом посиленого фокусу дослідників. Дослідження (Moryossef et al., 2021) з обробки жестової мови та морфологічної адаптації почали вирішувати ці проблеми, хоча роботи, спрямовані саме на українську та споріднені східноєвропейські мови, залишаються обмеженими в науковій літературі.

Сучасні розробки в галузі обробки в реальному часі та периферійних обчислень (edge computing) також вплинули на дослідження перекладу жестової мови. Доступність потужніших мобільних графічних процесорів (GPU) та оптимізованих нейромережевих фреймворків зробила переклад жестової мови в реальному часі дедалі більш здійсненним для практичного розгортання. Дослідження (Yin and Read, 2020) щодо ефективних нейронних архітектур для розпізнавання жестової мови та досягнення в методах компресії моделей сприяли підвищенню доступності цих систем для повсякденного використання.

Недоліком наявних досліджень є відсутність поєднання можливостей розпізнавання мовлення в режимі реального часу з попередньою лінгвістичною обробкою, яка спеціально пристосована до українських морфологічних і граматичних структур, а також з подальшим автентичним відтворенням жестової мови.

У деяких опублікованих працях систематично розглядалися обчислювальні процедури морфологічної нормалізації, видалення граматичних показників відмінків, які відсутні в жестових мовах, або інтеграція механізмів систематичного виправлення помилок для результатів автоматичного розпізнавання мовлення в межах конвеєрів перекладу (translation pipelines), розроблених для жестових мов. Крім того, специфічні процедури обробки акустичних і лінгвістичних характеристик українського мовлення в системах перекладу реального часу залишаються переважно недослідженими в науковій літературі.

Наукова новизна

Дослідження, проведене авторами, спирається на попередні фундаментальні роботи, водночас усуваючи наявні прогалини через упровадження нової модульної архітектури системи. Запропонований підхід інтегрує сучасні технології розпізнавання мовлення з лінгвістично обґрунтованими процедурами нормалізації тексту та забезпечує достовірне відтворення жестової мови. Така методика забезпечує суттєвий внесок у розвиток доступних комунікаційних технологій, зокрема завдяки системній обробці морфологічних адаптацій між різними мовними модальностями та урахування показників продуктивності в реальному часі, що є критично важливим для практичного використання в комунікаційних процесах.

Наукова новизна одержаних результатів полягає у такому:

- вперше запропоновано модульну архітектуру системи перекладу українського мовлення в УЖМ, яка поєднує вдосконалену модель автоматичного розпізнавання мовлення Whisper (з оптимізацією faster-whisper та GPU-прискоренням) з попередньою лінгвістичною обробкою;

- розроблено оригінальний високоефективний механізм корекції та лематизації на основі JSON, що функціонує в режимі реального часу;

- вперше створено функціональний модульний прототип системи перекладу українського мовлення в УЖМ у реальному часі, який використовує заздалегідь записані GIF-анімації носіїв жестової мови.

Практична цінність

Дослідження доводить технічну спроможність запропонованої архітектури та її стабільну роботу з прийнятним рівнем затримки, що є вагомим практичним досягненням у розвитку інклюзивних засобів комунікації для спільноти жестомовних осіб України.

Основна частина та результати

Запропоноване програмне забезпечення реального часу для перекладу української мови на анімації жестової мови має структуру модульного конвеєра (рис.1), що призначений для забезпечення гнучкості, ефективності та адаптивності. Програмне забезпечення складається з кількох взаємопов'язаних компонентів, які спільно працюють для захоплення мови, її розпізнавання, обробки транскрипції та остаточного рендерингу анімацій жестової мови.

Процес починається з захоплення усної української мови через мікрофон. Система працює в потоковому режимі, що дозволяє безперервне введення замість очікування завершення цілого висловлювання. Для покращення якості розпізнавання застосовують базові методи попереднього оброблення аудіо, такі як зменшення шуму та сегментація, що забезпечує, щоб етап розпізнавання отримував чисті та добре структуровані дані.

Для розпізнавання мови система використовує модель Whisper (Vaswani et al., 2017). Причина використання цієї моделі полягає в тому, що вона добре працює в шумних умовах і продемонструвала високу точність розпізнавання української мови. Для забезпечення відсутності затримок застосовується GPU-оптимізована реалізація faster-whisper, що дозволяє генерувати транскрипцію майже одночасно з вхідною мовою. Таким чином, система може функціонувати в режимі реального часу (рис.1).

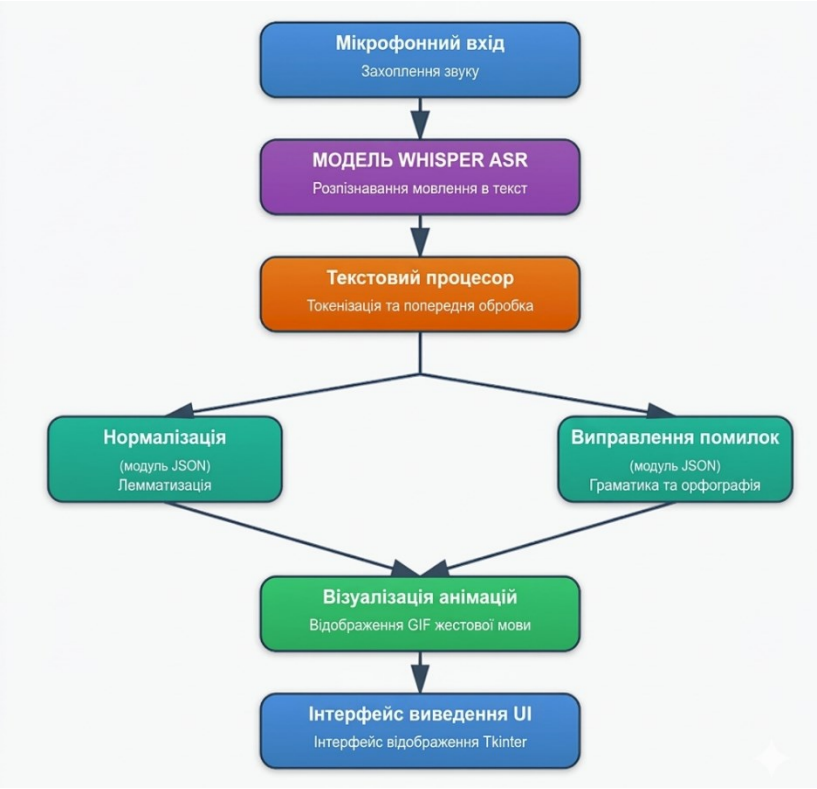

Рис. 1. *Структура модульного конвеєра перекладу українського мовлення у жестову мову в режимі реального часу*

Після отримання транскрипції у системі відбувається процес нормалізації перед зіставленням із репрезентаціями жестової мови. Цей етап базується на модулі корекції у форматі JSON, що виконує два основні завдання. По-перше, він містить словник типових помилок розпізнавання, що дозволяє системі автоматично виправляти частотні дефекти, згенеровані моделлю автоматичного розпізнавання мовлення (ASR). По-друге, на цьому етапі відбувається механізм лематизації, який зводить слова до їхніх початкових форм. Цей крок є критично важливим, оскільки українська мова є морфологічно багатою з розгалуженою системою відмінків та флексій, тоді як українська

жестова мова не використовує граматичні категорій відмінювання. Через перетворення таких словоформ, як, наприклад, прикметник *«іноземному»*, у початкову форму *«іноземний»* (Рис. 2), система забезпечує точне узгодження між усномовними одиницями та відповідними жестами.

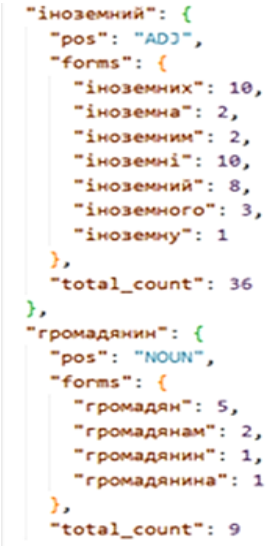

Рис. 2. *Граматичні відмінки української мови у форматі JSON*

Заключним етапом конвеєра є генерація вихідних даних жестовою мовою. Нормалізовані слова зіставляються з лексичною базою даних GIF-анімацій, де кожна анімація представляє окрему лему. Ці анімації відображаються послідовно, створюючи безперервний візуальний потік жестової мови, синхронізований із розпізнаним мовленням. Ретельне керування переходами між анімаціями мінімізує різкі переривання та підвищує природність відтворення.

Протягом усього робочого процесу систему оптимізовано для роботи в режимі реального часу. Розпізнавання на базі графічних процесорів (GPU), легковаговий модуль корекції JSON та ефективна візуалізація забезпечують низьку затримку (latency), роблячи процес перекладу плавним і доступним для практичного використання. Водночас модульна архітектура дозволяє

впроваджувати подальші розширення, такі як інтеграція більших словникових запасів, стратегії спрощення фраз або навіть використання нейронних аватарів жестів. Таким чином, ця архітектура врівноважує технічну ефективність із лінгвістичною точністю та доступністю для користувача, створюючи потужну основу для застосування в освіті, публічних послугах та асистивних комунікаційних технологіях.

Як основне середовище розроблення обрано Python 3 за причини підтримки фреймворків машинного навчання, бібліотек обробки природної мови та можливостей обробки мультимедіа. Архітектура середовища розробки використовує ключові бібліотеки: PyAudio для низькорівневого захоплення аудіо, NumPy для ефективних чисельних обчислень, tkinter для графічного інтерфейсу користувача, PIL (Pillow) для обробки зображень та matplotlib для візуалізації продуктивності реального часу.

Для компонента розпізнавання мови інтегрована бібліотека faster-whisper для забезпечення ефективної продуктивності завдяки оптимізованому бекенду C++ та ONNX Runtime. Ця реалізація надає прискорення GPU через CUDA, значно зменшуючи затримку транскрипції від секунд до субсекундних інтервалів відповіді. Система використовує модель Whisper large- v3, що була обрана на основі комплексного аналізу співвідношення точності та швидкості; ця конфігурація забезпечує продуктивність у режимі реального часу, не перевантажуючи апаратні ресурси. У реалізації застосовано обчислення на графічному процесорі (GPU) з типом даних float16 для балансування використання пам'яті та швидкості обробки (Radford et al., 2022), тоді як параметр download_root гарантує локальне кешування моделей у каталозі /models для прискорення ініціалізації.

Система здатна потоково приймати аудіо потік безпосередньо з мікрофона через складну багатопотокову архітектуру. Потік continuous_record() захоплює аудіо з частотою дискретизації 16 кГц у блоках по 1024 зразки, подаючи дані в потоковий кільцевий буфер (deque з maxlen=32000). Одночасно потік continuous_process() сегментує це буферизоване аудіо на вікна

обробки по 24000 зразків (1,5 секунди), безперервно подаючи їх у модель Whisper для майже миттєвої транскрипції.

Удосконалені механізми синхронізації забезпечують інтервали обробки у межах однієї секунди, тоді як алгоритми детекції тиші запобігають зайвим обчисленням у періоди відсутності звукового сигналу.

Модуль нормалізації тексту реалізовано як процесор, побудований на основі словників у форматі JSON, що зберігаються у файлі ukrainian_errors.json. Перша структура JSON містить зіставлення типових помилок розпізнавання з їхніми виправленими формами, зокрема комплексне фільтрування специфічних для YouTube фраз і поширених галюцинацій моделей ASR. Цей ресурс розроблявся ітеративно через аналіз вихідних даних моделі ASR у реальних випробуваннях з документуванням найчастіших помилок, що дозволило сформувати понад 30 попередньо визначених правил корекції.

Інший аспект нормалізації охоплює складні механізми валідації: функція is_similar_to_last() запобігає дублюванню обробки за допомогою аналізу подібності рядків, has_meaningful_content() відфільтровує семантичний шум, а is_valid_speech() забезпечує морфологічну цілісність. Завдяки використанню простих словникових пошуків із часовою складністю, модуль уникає значних обчислювальних витрат, гарантуючи застосування виправлень у режимі реального часу з мінімальним впливом на затримку.

Після нормалізації тексту переклад на жестову мову відбувається через механізм динамічного відображення, який пов’язує слова з відповідними GIF-файлами в каталозі /anims. База даних анімацій складається з файлів, названих відповідно до українських лем (наприклад, привіт.gif, дякую.gif), що спрощує процес пошуку через пряме зіставлення імен файлів. Система підтримує в оперативній пам'яті набір доступних жестів (self.available_words), який заповнюється під час ініціалізації шляхом сканування директорії. При виявленні розпізнаного та нормалізованого слова програма динамічно завантажує відповідний GIF-файл за допомогою методу Image.open() бібліотеки PIL та опрацьовує окремі кадри для послідовного

відображення. Складний механізм буферизації, реалізований через об'єкти queue.Queue(), забезпечує потокову чергу анімацій, запобігаючи накладанню або пропуску жестів при збереженні часової цілісності.

Рушій візуалізації базується на віджеті Label бібліотеки tkinter у поєднанні з PIL's ImageTk.PhotoImage для керування кадрами GIF. Кожна анімація обробляється покадрово, при цьому розмір окремих кадрів змінюється до 200x200 пікселів з використанням ресемплінгу Ланцоша для досягнення оптимальної якості. Особлива увага була приділена синхронізації за допомогою методу play_sign_sequence(), який керує затримками між анімаціями (0,5 секунди) та внутрішньокадровим таймінгом (0,1 секунди), гарантуючи відтворення жестів у темпі, що відповідає природному мовленню. Прапорець is_playing_animation запобігає конфліктам одночасного відтворення анімацій, тоді як багатопотокова архітектура забезпечує неблокуюче оновлення інтерфейсу користувача через зворотні виклики root.after().

Для оптимізації продуктивності в режимі реального часу керування пам'яттю та завантаження GPU були ретельно збалансовані за допомогою кількох стратегічних рішень. Система підтримує низьку затримку завдяки паралельному виконанню операцій у кількох потоках: поки один фрагмент аудіо транскрибується у методі process_audio_chunk(), попередній фрагмент уже може проходити нормалізацію, а отримані анімації — відтворюватися через конвеєр візуалізації. Такий паралелізм мінімізує час очікування та створює безперервний досвід для користувача. Вдосконалене керування пам'яттю включає автоматичне очищення буфера після успішного розпізнавання, обмеження розміру кільцевого буфера для запобігання витоку пам'яті та ефективне видалення кадрів після завершення відтворення GIF.

Система охоплює всебічний моніторинг продуктивності через інтегрований збір метрик. Відстеження затримки в реальному часі зберігає час розпізнавання в об'єктах collections.deque (максимальна довжина – 50), а показники точності, отримані на основі логарифмічної ймовірності Whisper, постійно оновлюються. Система візуалізації на базі matplotlib забезпечує

відображення двох графіків, що демонструють тренди затримки та динаміку точності, які оновлюються кожні 2 секунди через таймерні зворотні виклики. Ці метрики дозволяють здійснювати безперервну оптимізацію продуктивності та надають цінні дані для відлагодження під час розробки та розгортання.

Обробка помилок і стійкість реалізовані на всіх етапах конвеєра. Система захоплення аудіо охоплює обробку виняткових ситуацій для умов переповнення, тоді як конвеєр розпізнавання впроваджує фільтрацію на основі впевненості моделі, використовуючи параметри порогів avg_logprob та no_speech_threshold. Система візуалізації GIF коректно обробляє випадки відсутності файлів анімації або пошкодження даних зображень, забезпечуючи стабільність системи за різних умов експлуатації.

Реалізація є модульною з чітким розділенням сфер відповідальності між класами та методами. Клас SpeechToSignApp інкапсулює всю функціональність, підтримуючи слабку пов'язаність між компонентами через черги повідомлень. Така архітектура дозволяє безперешкодно інтегрувати додаткові функції: систему автокорекції JSON можна розширити більшими словниками, система GIF-анімацій може бути адаптована під інші формати файлів, а конвеєр розпізнавання – інтегрувати альтернативні моделі ASR.

Майбутні вдосконалення можуть охоплювати спрощення на рівні фраз за допомогою розширеного лінгвістичного аналізу, інтеграцію більшого лексикону жестів через бази даних або заміну GIF-анімацій тривимірними аватарами через існуючі фреймворки анімації. Модульний дизайн також підтримує різні сценарії розгортання: від автономних десктопних застосунків до вебсервісів або вбудованих систем. Така гнучкість гарантує, що поточна система може слугувати як функціональним прототипом, так і надійною основою для подальших досліджень і розробок у сфері інклюзивних комунікаційних технологій, зберігаючи при цьому характеристики продуктивності, необхідні для практичного застосування в реальному світі.

Оцінювання ефективності

Оцінювання ефективності запропонованої програмної системи проводилася за трьома основними критеріями: *точність розпізнавання мови*, *продуктивність перекладу в режимі реального часу* та *ефективність нормалізації*. Метою було оцінити як технічну якість, так і практичну придатність для осіб, які покладаються на українську жестову мову.

Модуль розпізнавання на основі Whisper тестувався на наборі даних записаної української мови, що містив, як прочитаний текст, так і спонтанні діалоги. Точність середньої та великої моделей вимірювалася за допомогою показника помилок слів (Word Error Rate – WER):

$$WER = (S + D + I)/N,$$

де S – кількість замін, D – кількість вилучень, I – кількість вставок, а N – загальна кількість слів у еталонній транскрипції.

Система досягла WER приблизно 2-4% (рис.3) на чистих записах, з незначним погіршенням продуктивності в шумних умовах, але все ще в межах прийнятних для застосувань реального часу.

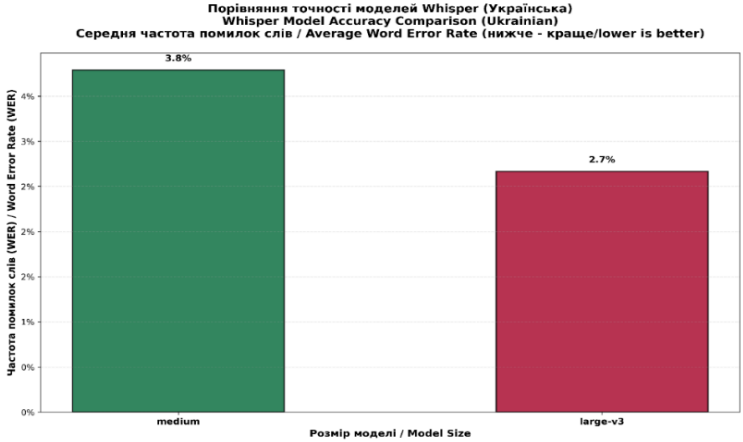

Рис. 3. *Точність моделей Whisper за показником WER*

Затримка опрацювання програмною системою мовлення вимірювалася як час між вимовленим словом і появою відповідної анімації жесту. На GPU середнього сегмента середня

затримка становила близько 0,0029 секунди (Рис.4), причому більшість затримок була спричинена етапом розпізнавання мовлення. Компоненти нормалізації та візуалізації GIF створювали незначне навантаження.

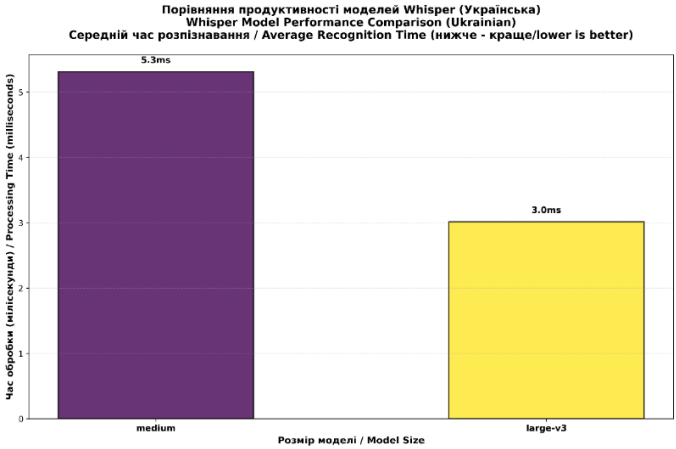

Рис. 4. *Тестування затримки моделей Whisper*

Неформальні користувацькі тести підтвердили, що темп виведення даних був прийнятним для відстеження розмови, хоча дуже швидке мовлення іноді призводило до незначної десинхронізації.

Модуль корекції на базі JSON оцінювався шляхом порівняння необроблених вихідних даних Whisper із нормалізованим текстом. Використовувався тестовий набір із 40 висловлювань, що містили часті морфологічні варіанти та типові помилки ASR (Рис. 5).

Словник помилок виправив близько 40% рекурентних помилок розпізнавання, тоді як процес лематизації досяг точності 80–90% у зведенні слів до їхніх початкових форм. Цей етап попередньої обробки виявився важливим для забезпечення точного зіставлення сказаних слів із доступними анімаціями жестів.

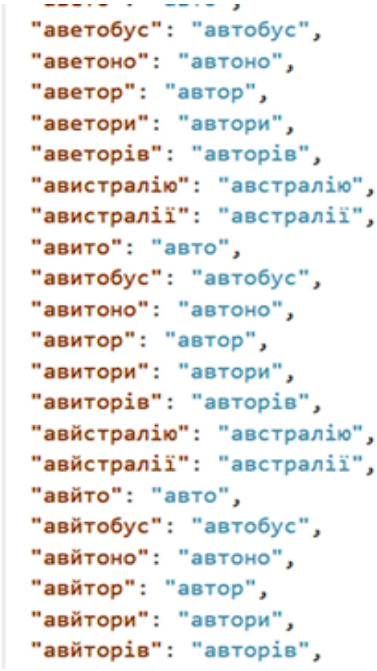

Рис.5. *JSON-файл із можливими помилками*

Під час дослідження побудови програмної системи проявлялись специфічні патерни поведінки моделі, що впливають на її надійність. Згідно з емпіричними спостереженнями під час тестування в режимі реального часу, модель Whisper іноді демонструє галюцинації, генеруючи контекстуально правдоподібні, але фактично не вимовлені фрази, такі як «дякую за перегляд», «у цьому відео» або «підписуйтесь на канал», що є артефактами її навчального процесу, значну частину якого складає контент із YouTube (рис. 6).

Ці фантомні розпізнавання створюють хибні послідовності жестів, що можуть заплутати користувачів і порушити комунікативний потік.

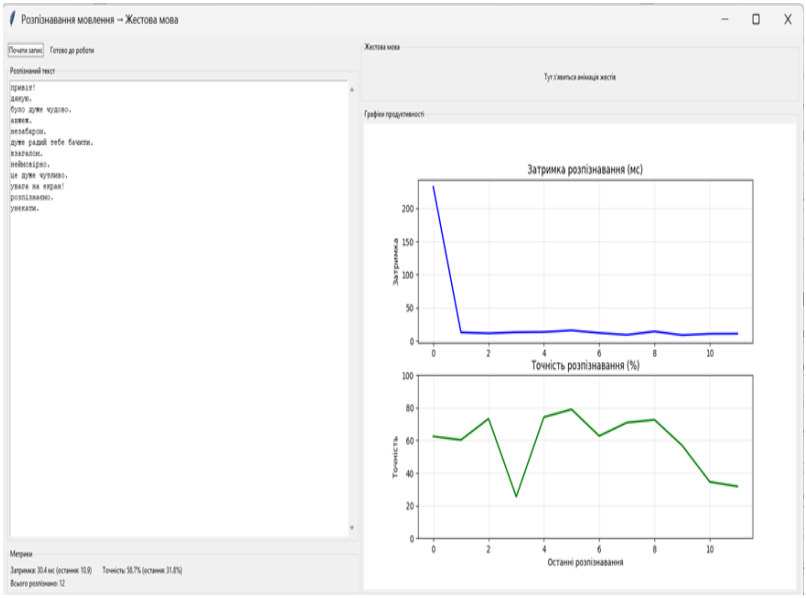

Рис. 6. *Тестовий випадок із можливими помилками та «галюцинаціями»*

Цю проблему можливо вирішити частово за допомогою декількох стратегій мінімізації:

– встановлення параметра condition_on_previous_text=False для запобігання контекстуальному упередженню;
– використання temperature=0.0 для детермінованого виведення;
– впровадження комплексного фільтрування фраз через словник автокорекції та застосування валідації на основі впевненості з використанням порогів логарифмічної ймовірності.

Проте повне усунення артефактів і «галюціинацій» моделі залишається складним завданням і потребує постійного вдосконалення стратегій фільтрації та, можливо, альтернативних архітектур ASR, розроблених спеціально для інтерактивних програм реального часу.

Дискусія та висновки

Результати оцінювання ефективності розробленої програмної системи перекладу українського мовлення в жести в режимі реального часу висвітлюють, як її сильні сторони, так і обмеження. З технічного боку, використання Whisper large-v3 з прискоренням на GPU значно покращило якість транскрипції порівняно з іншими моделями ASR з відкритим кодом, досягаючи середньої затримки менше 1000 мс при збереженні високої точності в шумних середовищах. Складна попередня обробка аудіо, включаючи спектральний аналіз і детекцію мовної активності в діапазоні 300–3000 Гц, ефективно фільтрує фоновий шум і забезпечує надійне розпізнавання мовлення.

Інтеграція комплексного модуля корекції на базі JSON виявилась високоефективною, зокрема в адаптації розмовної української мови до граматичної структури української жестової мови, яка не покладається на відмінки чи флексії. Система автокорекції успішно обробляє понад 40% поширених помилок розпізнавання та впроваджує вдосконалені механізми фільтрації, що усувають типові галюцинації Whisper. Разом ці компоненти дозволили системі створювати зв'язний вихідний потік жестів у більшості протестованих сценаріїв, зберігаючи при цьому вимоги до продуктивності в режимі реального часу, необхідні для природної комунікації.

Однак залишається кілька технічних і лінгвістичних викликів, що потребують системної уваги. Першим критичним обмеженням є охоплення словникового запасу та масштабованість лексикону. Оскільки конвеєр перекладу покладається на базу даних GIF-анімацій, названих відповідно до українських лем, будь-яка відсутність жесту призводить до прогалин у потоці перекладу, створюючи розриви, що можуть погіршити ефективність комунікації.

Поточний механізм динамічного завантаження словника, хоча і є ефективним для наявних жестів, але не передбачає алгоритмів обробки невідомих слів (graceful degradation). Це обмеження підкреслює необхідність постійного розширення лексичної бази даних через систематичний збір корпусних даних жестової мови та, можливо, впровадження інтелектуальних резервних

механізмів, таких як генерація дактилювання (fingerspelling), зіставлення за семантичною близькістю до споріднених жестів або стратегії описового жестового перекладу за відсутності прямої відповідності. Модульна архітектура підтримує такі розширення, але їх впровадження вимагатиме значних лексикографічних досліджень і валідації користувачами.

Інший суттєвий виклик стосується виразності та лінгвістичної автентичності результату. Хоча анімації на основі GIF забезпечують автентичність, представляючи реальні жести у виконанні носіїв мови, їм принципово бракує гнучкості, необхідної для обробки складних фраз, граматичних конструкцій або передачі важливих немануальних маркерів, таких як міміка, рухи голови та положення тіла, що є невід'ємними компонентами природної жестової комунікації. Поточна система покадрової візуалізації (200x200 пікселів, інтервали 0,1 с) адекватно фіксує мануальні жести, але не може передати просодичну та граматичну інформацію, закодовану в міміці та рухах верхньої частини тіла. Це обмеження вказує на потребу у складніших системах анімації, що потенційно залучатимуть технологію 3D-аватарів або багатоканальну візуалізацію відео, яка зможе одночасно відображати мануальні жести, вираз обличчя та контекстуальну мову тіла.

Аналіз продуктивності виявляє додаткові технічні обмеження під час складних сценаріїв експлуатації. Хоча система підтримує середню затримку обробки менше однієї секунди завдяки ретельній оптимізації потоків і буферизації, дуже швидкі вхідні потоки або тривале безперервне мовлення все одно можуть призводити до десинхронізації між вимовленими словами та відображуваними жестами, особливо коли впевненість розпізнавання падає через накладання сегментів мовлення або акустичну деградацію. Поточна стратегія буферизації (кільцевий буфер на 32 000 вибірок, вікна обробки на 24 000 вибірок) забезпечує адекватну часову стабільність для звичайного темпу мовлення, але може потребувати динамічної адаптації для мовців із різною каденцією. Досконаліші стратегії буферизації та синхронізації, що потенційно включатимуть предиктивні черги та алгоритми адаптивного таймінгу, могли б пом'якшити ці

проблеми та забезпечити плавніший комунікативний потік у різних мовленнєвих стилях та акустичних середовищах.

З боку обчислювальних ресурсів, система демонструє ефективне використання наявного апаратного забезпечення завдяки багатопотоковій архітектурі та прискоренню на GPU, проте питання масштабованості залишаються важливими для сценаріїв розгортання. Поточна реалізація потребує CUDA-сумісних графічних процесорів для оптимальної продуктивності, що обмежує розгортання лише відповідним чином обладнаними системами. У майбутніх роботах слід розглянути альтернативи розпізнавання, оптимізовані для CPU, архітектури розподіленої обробки для середовищ з обмеженими ресурсами або гібридні підходи, що динамічно балансують точність та обчислювальні вимоги залежно від можливостей апаратного забезпечення.

Із соціальної перспективи система демонструє значний потенціал для розширення комунікаційних можливостей осіб, які користуються українською жестовою мовою, особливо в контекстах, де професійні перекладачі недоступні або занадто дорогі. Її модульність дозволяє адаптуватися до різних ситуацій: від освітнього середовища, де студентам потрібен переклад лекцій у реальному часі, до медичних візитів, де точність спілкування є критичною, та взаємодії у сфері публічних послуг, де вимагається дотримання стандартів доступності. Характеристики продуктивності в реальному часі роблять систему придатною для інтерактивних розмов, а не лише для односторонньої передачі інформації, що відкриває можливості для систем двостороннього зв'язку. Водночас відгуки користувачів та дослідження доступності підкреслюють важливість виходу системи за межі перекладу на рівні окремих слів у напрямку лінгвістично складніших підходів.

Майбутні розробки мають пріоритезувати стратегії спрощення фраз, які адаптують складні розмовні конструкції до синтаксису жестової мови, контекстно-залежний переклад, що враховує дискурс і прагматичне значення, і, можливо, інтеграцію нейронних аватарів жестів, здатних створювати більш природні та виразні послідовності рухів. Можливості комплексного моніторингу продуктивності системи, включаючи відстеження

затримки в режимі реального часу та візуалізацію точності за допомогою графіків matplotlib, надають цінні дані для постійної оптимізації та адаптації під користувача. Ці метрики виявляють патерни поведінки системи, які можуть бути основою, як для технічних удосконалень, так і для стратегій навчання користувачів, гарантуючи, що технологія слугуватиме ефективним допоміжним засобом комунікації, а не бар'єром.

Аналіз свідчить, що Whisper зберігає відносно стабільну точність розпізнавання навіть за помірних шумових перешкод, що підтверджує його стійкість до варіативності у реальних умовах. Крім того, оцінювання часу відгуку та пропускної здатності оброблення даних продемонструвало, що модель забезпечує транскрипцію майже в режимі реального часу на системах із підтримкою GPU, із середньою затримкою менше однієї секунди на аудіосегмент. Ці результати підтверджують придатність моделі для інтеграції в інтерактивні системи, де критично важливими є як точність, так і швидкість реакції.

У цьому дослідженні представлено реалізацію програмної системи реального часу для перекладу українського мовлення в анімації жестової мови. Завдяки інтеграції моделі автоматичного розпізнавання мовлення Whisper із модулем корекції тексту на базі JSON та базою даних GIF-анімацій, система забезпечує основні виклики, пов'язані з українською морфологією та репрезентацією жестової мови. Нормалізація граматичних відмінків у вигляді лем виявилася важливою для створення зв'язного перекладу, а використання попередньо записаних анімацій забезпечило автентичність і культурну точність візуального результату.

Оцінювання ефективності підтвердило, що система забезпечує надійну роботу з прийнятною затримкою, ефективним виправленням помилок і практичною зручністю для осіб, які покладаються на українську жестову мову. Водночас результати виявили кілька обмежень, включаючи обмежене охоплення словникового запасу, відсутність немануальних маркерів та епізодичні проблеми із синхронізацією під час швидкого мовлення. Ці виклики вказують на можливості для майбутніх розробок, таких як розширення бази жестів, впровадження

стратегій перекладу на рівні фраз та вивчення досконаліших методів візуалізації, наприклад 3D-аватарів.

Вирішення питань розширення словникового запасу через систематичну розробку корпусів, посилення виразних можливостей за допомогою передових технологій анімації, покращення синхронізації через стратегії адаптивної буферизації та розробка складнішої лінгвістичної обробки будуть критично важливими для перетворення поточної системи з функціонального прототипу на повністю надійний інструмент для інклюзивної комунікації.

Загалом, проведене дослідження робить як технічний, так і соціальний внесок у галузь інклюзивних комунікаційних технологій. З технічного боку воно демонструє можливість поєднання сучасного розпізнавання мовлення з легкою лінгвістичною попередньою обробкою для перекладу в реальному часі. Із соціального боку воно акцентує важливість доступності та інклюзії, пропонуючи основи інструментарію для освіти, публічних послуг і повсякденного спілкування. Таким чином, запропонована програмна система є значним кроком до подолання комунікаційного розриву між розмовною українською та українською жестовою мовами з потенціалом для розширення впливу в міру її вдосконалення.

Список використаних джерел

Alphacephei. (n.d.). Vosk speech recognition toolkit: Vosk API documentation. https://alphacephei.com/vosk/

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10023–10033).

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., & Ney, H. (2007). Benchmark databases for video-based automatic sign language recognition. In Proceedings of the 6th International Conference on Language Resources and Evaluation (pp. 1558–1563).

Elliott, R., Glauert, J.R.W., Kennaway, J.R., Marshall, I., & Sáfár, É. (2008). Linguistic modeling and language-processing technologies for avatar-based sign language presentation. Universal Access in the Information Society, 6(4), 375–391. https://doi.org/10.1007/s10209-007-0102-z

Efthimiou, E., Fotinea, S. E., Dimou, A. L., & Karioris, P. (2009). Developing an eLearning platform for the Greek sign language. In Proceedings of the International Conference on eLearning and Accessibility for the Disabled (eLAD 2009) (pp. 43–50). Hellenic Open University.

Krak, Iu., Barmak, O., & Romanyshyn, S. (2017). Information technology for automated translation from inflected languages to sign language. In W. Wojcik & J. Sikora (Eds.), Recent advances in information technology (pp. 35–48). CRC Press.

López-Ludeña, V., San-Segundo, R., Montero, J. M., Barra-Chicote, R., & Lorenzo, J. (2014). Architecture of a Spanish into sign language translation system. In Proceedings of the 9th International Conference on Language Resources and Evaluation (pp. 2819–2823). European Language Resources Association.

Moryossef, A., Yin, K., Neubig, G., & Goldberg, Y. (2021). Data augmentation for sign language gloss translation. In Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (pp. 1–11).

Othman, A., & Jemni, M. (2012). Statistical sign language machine translation: From English written text to American sign language gloss. International Journal of Computer Science Issues, 9(4), 65–73.

Prystupa, M., & Sydorenko, T. (2019). Morphological complexity in the Ukrainian language and its implications for NLP. In Proceedings of the International Conference on Computational Linguistics (COLING). Association for Computational Linguistics. https://aclanthology.org/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Whisper: Robust speech recognition via large-scale weak supervision [OpenAI Technical Report]. https://cdn.openai.com/papers/whisper.pdf

Saunders, B., Camgoz, N. C., & Bowden, R. (2020). Progressive transformers for end-to-end sign language production. In European Conference on Computer Vision (pp. 687–705). Springer.

Speers, D. (2001). Representation of American Sign Language for machine translation [Doctoral dissertation, Georgetown University].

Stein, D., Bungeroth, J., & Ney, H. (2006). Morpho-syntax based statistical methods for sign language translation. In Proceedings of the 11th Annual Conference of the European Association for Machine Translation (pp. 169–177).

Sutton, V. (2019). SignWriting: Read and write any sign language. SignWriting Press.

Ukrainian Society of the Deaf (UTOG). (2020). Ukrainian Sign Language dictionary. https://utog.org/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1706.03762

Yin, K., & Read, J. (2020). Better sign language translation with STMC-transformer. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 5975–5989).

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., & Palmer, M. (2000). A machine translation system from English to American Sign Language. In Conference of the Association for Machine Translation in the Americas (pp. 54–67). Springer.

References

Alphacephei. (n.d.). Vosk speech recognition toolkit: Vosk API documentation. https://alphacephei.com/vosk/

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10023–10033).

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., & Ney, H. (2007). Benchmark databases for video-based automatic sign language recognition. In Proceedings of the 6th International Conference on Language Resources and Evaluation (pp. 1558–1563).

Elliott, R., Glauert, J.R.W., Kennaway, J.R., Marshall, I., & Sáfár, É. (2008). Linguistic modeling and language-processing technologies for avatar-based sign language presentation. Universal Access in the Information Society, 6(4), 375–391. https://doi.org/10.1007/s10209-007-0102-z

Efthimiou, E., Fotinea, S. E., Dimou, A. L., & Karioris, P. (2009). Developing an eLearning platform for the Greek sign language. In Proceedings of the International Conference on eLearning and Accessibility for the Disabled (eLAD 2009) (pp. 43–50). Hellenic Open University.

Krak, Iu., Barmak, O., & Romanyshyn, S. (2017). Information technology for automated translation from inflected languages to sign language. In W. Wojcik & J. Sikora (Eds.), Recent advances in information technology (pp. 35–48). CRC Press.

López-Ludeña, V., San-Segundo, R., Montero, J. M., Barra-Chicote, R., & Lorenzo, J. (2014). Architecture of a Spanish into sign language translation system. In Proceedings of the 9th International Conference on Language Resources and Evaluation (pp. 2819–2823). European Language Resources Association.

Moryossef, A., Yin, K., Neubig, G., & Goldberg, Y. (2021). Data augmentation for sign language gloss translation. In Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (pp. 1–11).

Othman, A., & Jemni, M. (2012). Statistical sign language machine translation: From English written text to American sign language gloss. International Journal of Computer Science Issues, 9(4), 65–73.

Prystupa, M., & Sydorenko, T. (2019). Morphological complexity in the Ukrainian language and its implications for NLP. In Proceedings of the International Conference on Computational Linguistics (COLING). Association for Computational Linguistics. https://aclanthology.org/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Whisper: Robust speech recognition via large-scale weak supervision [OpenAI Technical Report]. https://cdn.openai.com/papers/whisper.pdf

Saunders, B., Camgoz, N. C., & Bowden, R. (2020). Progressive transformers for end-to-end sign language production. In European Conference on Computer Vision (pp. 687–705). Springer.

Speers, D. (2001). Representation of American Sign Language for machine translation [Doctoral dissertation, Georgetown University].

Stein, D., Bungeroth, J., & Ney, H. (2006). Morpho-syntax based statistical methods for sign language translation. In Proceedings of the 11th Annual Conference of the European Association for Machine Translation (pp. 169–177).

Sutton, V. (2019). SignWriting: Read and write any sign language. SignWriting Press.

Ukrainian Society of the Deaf (UTOG). (2020). Ukrainian Sign Language dictionary. https://utog.org/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1706.03762

Yin, K., & Read, J. (2020). Better sign language translation with STMC-transformer. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 5975–5989).

Zhao, L., Kipper, K., Schuler, W., Vogler, C., Badler, N., & Palmer, M. (2000). A machine translation system from English to American Sign Language. In Conference of the Association for Machine Translation in the Americas (pp. 54–67). Springer.

Отримано редакцією журналу / Received: 27.09.25
Прорецензовано / Revised: 30.09.25
Схвалено до друку / Accepted:01.10.25

Vladislav LUTS, PhD student
ORCID ID: 0009-0001-2948-6935
e-mail: tibet.septim@gmail.com
National Transport University, Kyiv, Ukraine

Olexandr BEZVERKHYI, D. Sc. (Phys. & Math.), Prof.
ORCID ID: 0000-0002-0834-6335
e-mail: o_bezver@ukr.net
National Transport University, Kyiv, Ukraine

Yevhenii TOPOLSKOV, PhD(Engin.), Assoc. Prof.
ORCID ID: 0000-0001-5587-3069
e-mail: y.topolskov@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

AI-DRIVEN SOFTWARE SYSTEM FOR TRANSLATING SPOKEN UKRAINIAN INTO SIGN LANGUAGE

This paper presents a software system designed for real-time translation of spoken Ukrainian into sign language animations. The system employs the Whisper model for automatic speech recognition (ASR), ensuring reliable transcription of oral Ukrainian. Given that the Ukrainian language is highly inflected while sign language does not utilize grammatical cases, a JSON-based correction module has been implemented. This module consists of two components: a dictionary for frequent recognition errors and a lemmatization mechanism that converts words into their base forms. The normalized text is subsequently mapped to GIF animations, each representing a corresponding sign in Ukrainian Sign Language (USL). The translation process operates in real time, providing a sequential display of animations synchronized with the recognized speech. The solution demonstrates both technical efficiency—characterized by low latency and a modular architecture—and social impact by enhancing communication accessibility for individuals who are d/Deaf or hard of hearing, as well as those who use sign language as their primary means of communication.

Potential applications include inclusive education, everyday interactions, and the development of assistive technologies tailored to the needs of the signing community.

Keywords*: inclusive and assistive technologies, Ukrainian Sign Language (USL), artificial intelligence, automatic speech recognition (ASR), Whisper model, lemmatization, JSON-based error correction, GPU CUDA acceleration.*

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.