https://doi.org/10.15407/csc.2024.04.039
УДК 658.512

Є.Р. МРОЗЕК, аспірант, відділ Розпізнавання та синтезу звукових образів,
Міжнародний науково-навчальний центр інформаційних технологій і систем НАН та МОН України,
просп. Академіка Глушкова, 40, м. Київ, Україна, 03187,
ORCID: https://orcid.org/0009-0008-4989-5016,
просп. Академіка Глушкова, 40, м. Київ, Україна, 03187,
zekamrozek@gmail.com

АНАЛІЗ СУЧАСНИХ ПІДХОДІВ ДО РОЗВ'ЯЗАННЯ ЗАДАЧ РОЗПІЗНАВАННЯ МОВЛЕННЯ

Необхідність сучасних підходів до розв'язання задач розпізнавання мови зумовлена швидким розвитком штучного інтелекту та необхідністю покращення точності й швидкості взаємодії людини з комп'ютером у різних сферах, таких як голосові помічники, переклад та автоматизація. Цей напрям стає дедалі актуальнішим через зростання обсягів згенерованих аудіоданих та необхідності їхньої обробки в реальному часі, зокрема в українських реаліях, де поєднуються кілька мов та діалектів. На цей час існує кілька підходів до розпізнавання, аналізу та транскрибування мовлення, зокрема методи на базі нейронних мереж, методи діаризації співрозмовників, видалення шуму та структуризації даних. Проте залишається актуальною проблема створення універсального рішення, яке б відповідало потребам багатомовних середовищ і дозволяло ефективно працювати з неструктурованими аудіоданими.

Метою статті є огляд наявних інструментів та алгоритмів для розв'язання задачі розпізнавання мови, зокрема української. Використовуються методи розпізнавання мови, глибоке навчання, трансформери. Для побудови бази даних і знань системи багатомовного усного діалогу було розглянуто теоретичне підґрунтя підходів та моделей для розпізнавання мови. Також досліджено ефективні приклади покращення точності транскрибування для мов з обмеженими даними та потенційні кроки збільшення швидкодії системи. Розглянуто потенційні дані для навчання моделей, наведено структурований огляд сучасних методів обробки та аналізу багатомовних аудіофайлів, їхніх переваг та недоліків, а також визначення невирішених проблем.

Ключові слова: *розпізнавання мови, нейронні мережі, машинне навчання, багатомовний усний діалог.*

Вступ

Автоматичне розпізнавання мовлення — це сфера, де поєднуються досягнення комп'ютерного зору (*CV*), обробки природної мови (*NLP*) та специфічні аспекти аудіообробки, як от зниження рівня шуму, покращення якості звуку, виокремлення мовних сигналів у складних акустичних середовищах. Інтеграція цих технологій дозволяє створювати системи, здатні не лише розпізнавати мову, а й аналізувати контекст, інтонацію та інші важливі аспекти комунікації. Ця технологія стає дедалі важливішою в сучасному світі, оскільки дає змогу ефективно аналізувати аудіодані та перетворювати їх на структуровану інформацію. Це особливо актуально

Cite: Мрозек Є.Р. Аналіз сучасних підходів до розв'язання задач розпізнавання мовлення. *Control Systems and Computers*, 2024, 4, 39—49. https://doi.org/10.15407/csc.2024.04.039

для країн, де мовний ландшафт складається з кількох мов і діалектів, як, наприклад, в Україні.

Системи автоматичного розпізнавання мовлення знаходять своє застосування в багатьох сферах: від автоматизації кол-центрів та телефонних служб до створення інтелектуальних ботів, які здатні обробляти багатомовні діалоги в режимі реального часу. Попит на подібні рішення постійно зростає, адже мовлення є природним та швидким способом передачі інформації.

Попри досягнутий прогрес у галузі чимало проблем залишаються нерозв'язаними. Наприклад, сучасним алгоритмам розпізнавання мовлення часто бракує точності в умовах шуму або багатомовності, а системи, що базуються на нейронних мережах, потребують великих ресурсів для навчання. Тому актуальними напрямами розвитку є: підвищення точності розпізнавання, адаптація систем до обмежених ресурсів і покращення роботи в режимі реального часу.

У статті аналізуються основні підходи до автоматичного розпізнавання мовлення, їхні переваги та недоліки, а також розглядаються можливості застосування таких технологій у реаліях українського ринку.

Перші спроби вирішення задачі розпізнавання мови

Метод прихованих марковських моделей [1] (*HMM*) став першим значним проривом в автоматичному розпізнаванні мовлення ще у 1970—1980-х роках. Це був один з перших підходів, який надавав структуровану модель для аналізу звукових сигналів, представляючи мовлення як послідовність фонем, кожна з яких перебувала в певному стані. *HMM* дозволяли моделювати ймовірності переходів між звуками, що робило можливим точне розпізнавання слів і фраз.

Гаусові змішані моделі (*GMM*) [2] було додано до *HMM* у 1990-х роках для покращення моделювання звукових характеристик. *GMM* уможливили кращий опис різноманітності звукових сигналів, представляючи кожну фонему як суміш гаусових розподілів. Це дало змогу підвищити точність розпізнавання мовлення, зробивши системи гнучкішими щодо змін інтонації та вимови, але вони мали кілька ключових проблем, що обмежували їхню ефективність. Розглянемо їх далі.

1. Незалежність від контексту: *HMM* не враховували, як звуки взаємопов'язані в мовленні, тому часто не могли правильно розпізнати слова.

2. Проблеми з варіаціями звуків: *GMM* погано справлялися з різними вимовами або фоновим шумом, тому в складних умовах точність розпізнавання знижувалася.

3. Необхідність великої кількості даних: щоб ці моделі працювали добре, їм потрібно було багато прикладів мовлення.

Звісно *HMM* та *GMM* потребують набагато менших даних, ніж сучасні системи, але вони мали обмежену здатність опрацьовувати складні мовні варіації. Тому навіть із великою кількістю даних їхня продуктивність швидко досягала межі. Трансформери ж можуть ефективно використовувати величезні обсяги даних і при цьому забезпечують набагато кращі результати, особливо в умовах шуму та багатомовності.

Середня похибка за *WER* цих підходів 15—20 % на академічно чистих корпусах, як *Wall Street Journal* та *Switchboard*.

Підходи з використанням нейронних мереж

Наступним великим прогресом у розпізнаванні мовлення була популяризація глибоких нейронних мереж (*DNN*) [3], які стали ключовим інструментом сучасних систем. *DNN* дозволили моделювати складні зв'язки між аудіосигналами та текстом, вивчаючи важливі характеристики на різних рівнях. Це значно підвищило точність порівняно з традиційними підходами.

Спектрограми почали широко використовуватися з появою *DNN* у 2010-х роках, коли їхні можливості для обробки та аналізу складних сигналів було визнано особливо корисними для розпізнавання мовлення. Раніше аудіосигнали оброблялися в часовій області, що не дозволяло вповні використовувати частотні ознаки. Спектрограми ж дозволили перевести звукові сигнали в графічне представлення частотної

інформації, що значно полегшило роботу нейронних мереж. Спочатку для обробки спектрограм застосовувалися конволюційні нейронні мережі (*CNN*), створені для аналізу зображень, але вони виявилися дуже ефективними для вивчення патернів у спектрограмах. Це стало ключовим кроком у підвищенні точності та надійності систем розпізнавання мовлення.

Середня похибка *WER* моделей з нейронними підходами коливається від 5—10 % на академічних корпусах.

Більшість методів автоматичного розпізнавання мовлення, таких як *HMM*, *GMM*, *DNN*, *CNN*, *TDNN* і навіть *RNN* (особливу роль відіграли *LSTM*), здебільшого фокусувалися на англійській мові та інших поширених мовах, таких як китайська, іспанська, німецька. Українська мова не була пріоритетною для великих комерційних або академічних систем розпізнавання мовлення, оскільки для неї не було достатньо великих корпусів даних, необхідних для навчання моделей. Хоча деякі моделі, як-от *Kaldi*, можна було адаптувати для української мови, більшість систем, особливо комерційних, мали обмежену підтримку української мови.

Трансформери для задачі розпізнавання мови

Наступним кроком стали трансформери — це тип нейронних мереж, що були вперше представлені в 2017 році в статті «*Attention is All You Need*» [4] і є основою багатьох сучасних моделей обробки природної мови та мовлення. Головною особливістю трансформерів є використання механізму уваги (*attention*), зокрема самоуваги (*self-attention*), який дозволяє моделі ефективно фокусуватися на різних частинах вхідної послідовності, враховуючи контекст для кожного елемента. Особливості цього підходу є такими:

1. Блок *Attention* допомагає краще навчатися на довгих аудіо завдяки запам'ятовуванню інформації з довшого аудіофрагмента.

2. Також можливість навчатися різними мовами одночасно. Це дає змогу узагальнювати знання між різними мовами.

3. Також механізм *self-attention* дає змогу краще враховувати контекст.

4. Менша залежність від чистоти аудіо. Адже трансформери демонструють кращу стійкість до шуму та фонових перешкод порівняно з попередніми моделями

5. Можливість вирішувати складніші задачі, такі як онлайн-переклад.

В таблиці 1 можна побачити похибки *WER* різних моделей на корпусах ближчих до реального мовлення.

Як бачимо, похибка *WER* транформерів є набагато меншою, ніж у попередників.

Особливу увагу хочеться звернути на *Whisper* [5]. *Whisper* — це модель для розпізнавання мовлення, розроблена компанією *OpenAI* і випущена у 2022 році. Модель вирізняється високою точністю розпізнавання мовлення навіть у складних акустичних умовах, таких як наявність шуму або низька якість запису. Багато моделей навчалися на академічних корпусах зібраних даних, але вони не мають такої ж робастності як Whisper на інших корпусах без донавчання. На рис. 1 можна побачити залежності похибки *WER* різних підходів для академічно чистого корпусу *LibriSpeech* та близького до реального життя *Common Voice*.

Whisper є потужною багатомовною моделлю, що здатна розпізнавати багато мов, але вона працює з одною мовою за сеанс розпізнавання. Одним зі способів розпізнавання кількох мов у рамках одного речення є підхід *Code-Switching*, який потребує подальших досліджень. Одна з

Таблиця 1. **Порівняння точності моделей з використанням трансформерів**

Назва моделі	Рік	*WER* (%)	Перелік корпусів
Wav2Vec 2.0	2020	3,2	*LibriSpeech, CommonVoice*
Wav2Vec2-BERT	2021	2,8	*LibriSpeech, CommonVoice*
Data2Vec	2022	4,3	*LibriSpeech*
Whisper	2022	4,1	*LibriSpeech, Multilingual datasets*
Squeezeformer	2022	5,1	*LibriSpeech*

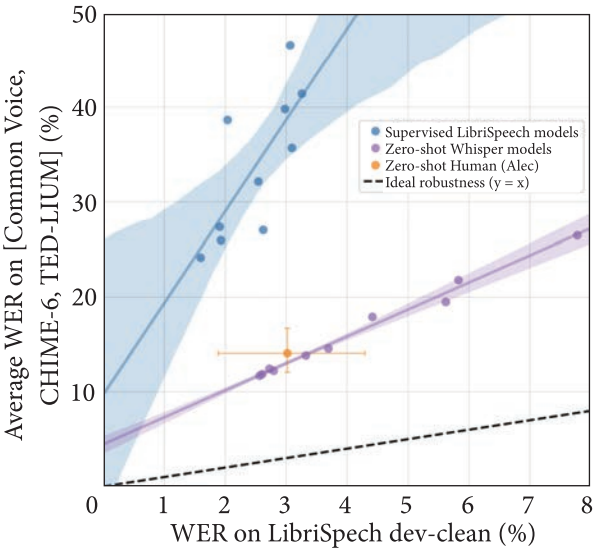

Рис. 1. Графіки порівняння точності моделей на корпусі *LibriSpeech* та *Common Voice*, *CHiME*-6, *TED-LIUM*

ключових особливостей *Whisper* — це її відкриті ваги, що дозволяє дослідникам та розробникам вільно використовувати модель для своїх проєктів, а також адаптувати її для різних мов, включно з українською.

Як можна побачити на рис. 2, що в архітектурі *Whisper* використано найсучасніші блоки, що вже добре зарекомендували себе в розробці моделей для розпізнавання мовлення та обробки тексту. *Whisper* використовує трансформери, які широко застосовуються в багатьох моделях для аналізу послідовностей, але в самій архітектурі нічого кардинально нового не було створено.

Головним внеском *Whisper* вкотре стало підтвердження теорії, що велика кількість даних відіграє ключову роль у створенні моделей, які є робастними, точними та здатні узагальнювати знання для різних мов і акустичних умов. *Whisper* була натренована на величезних обсягах багатомовного аудіо, що дало змогу досягти високої продуктивності навіть у шумних середовищах і на мовах з обмеженими даними, як-от українська. Це ще раз підтверджує, що якість і кількість даних є критичними факторами для успішної роботи моделей розпізнавання мовлення.

Whisper від *OpenAI* була натренована на величезному наборі даних, що включав й анотовані, й неанотовані аудіо-записи. Основна частина даних — це багатомовні аудіофайли, що охоплюють широкий спектр мов і акустичних умов.

Підготовка корпусу для *Whisper*:

• Кількість аудіо-годин: *Whisper* була навчена на 680000 годин аудіо, що становить одну з найбільших колекцій даних для навчання моделі розпізнавання мовлення.

• Дані з анотаціями: із цих 680000 годин приблизно 117000 годин є анотованими, тобто вони містять транскрипції, що використовуються для навчання моделей зі спостереженням (*supervised learning*). Ці дані охоплюють велику кількість мов, що дозволяє моделі добре узагальнювати мовні ознаки і працювати у багатомовних середовищах.

• Неанотовані дані: решта даних (~ 563000 годин) не мають текстових транскрипцій. Вони використовуються в процесі самонавчання моделі (*self-supervised learning*), що дозволяє їй вивчати загальні ознаки аудіо навіть без безпосереднього зіставлення з текстом. Це допомагає моделі навчатися на величезних обсягах даних, що підвищує її робастність і точність.

Особливості корпусу:

• Мультиковість: *Whisper* була навчена на записах понад 50 мов, що уможливлює розпізнавання мовлення різними мовами та діалектами, включно із менш популярними мовами.

• Різноманітність умов: корпус містив записи в різних акустичних умовах — від чистих студійних записів до шумних аудіо, що зробило модель дуже стійкою до шумів та інших зовнішніх перешкод.

• Різні типи мовлення: включає аудіо з різних джерел, таких як телефонні розмови, лекції, публічні виступи та інші типи мовлення.

На графіку (Рис. 3) показано кореляцію між кількістю транскрибованого аудіо та продуктивністю моделі розпізнавання мовлення (*Word Error Rate*, *WER*) для різних мов.

Графік підтверджує, що чим більше транскрибованого аудіо є доступним для певної мови, тим кращою є продуктивність моделі розпіз-

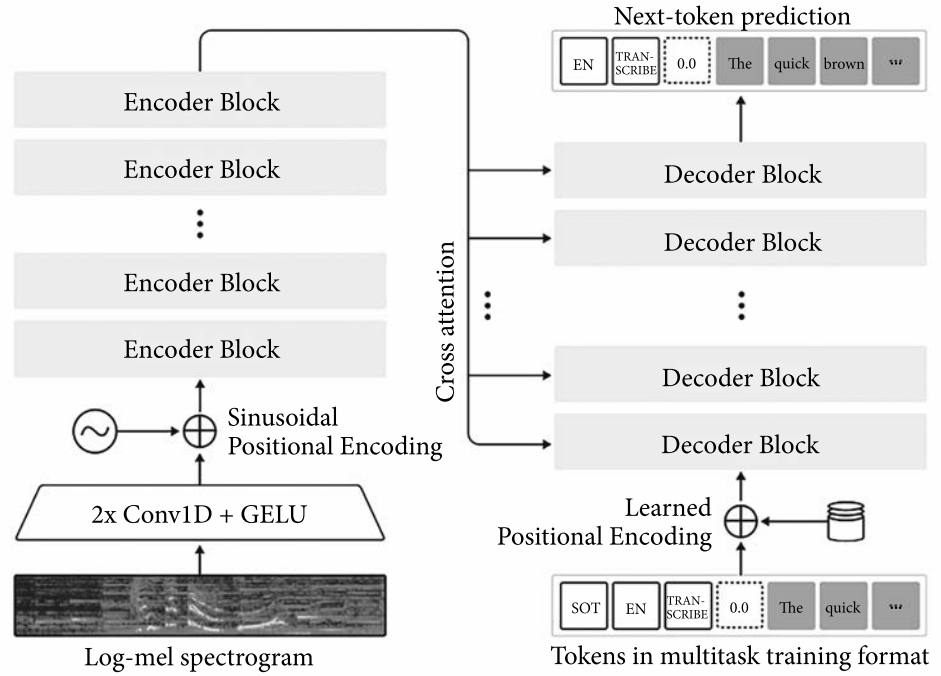

Рис. 2. Архітектура *Whisper*

навання мовлення для цієї мови (нижчий *WER*). Це також свідчить про те, що великі набори даних є критичними для досягнення високої точності в задачах розпізнавання мовлення, особливо для мов із обмеженими ресурсами

Загальні проблеми із обробкою аудіо

Обробка аудіо для розпізнавання мовлення стикається з низкою проблем, які можуть значно вплинути на точність роботи моделей. Найпоширенішими викликами є шум, низька якість аудіо, а також варіативність голосів різних спікерів. Ці фактори можуть суттєво ускладнити процес розпізнавання мовлення, особливо тоді, коли система працює в реальному часі або у складних акустичних умовах.

Шум — це одна з найбільших проблем для будь-якої системи розпізнавання мовлення. Шум може надходити з різних джерел: вуличні звуки, фонові розмови, звуки техніки та інше. Проблема полягає в тому, що модель часто не може відрізнити важливу інформацію (голос

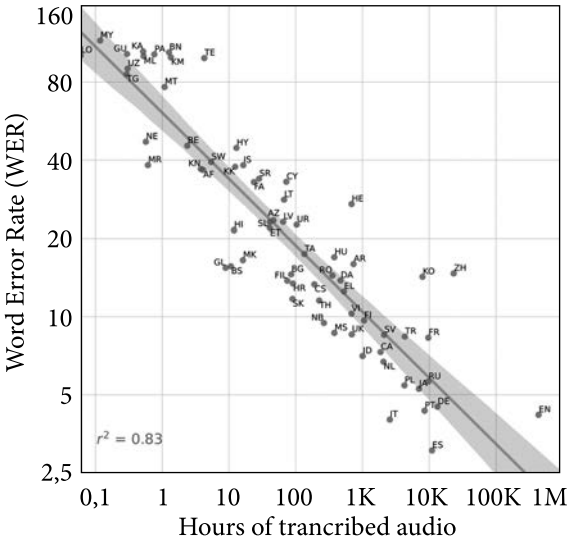

Рис. 3. Залежність точності від кількості анотованих даних

мовця) від сторонніх шумів, що може призвести до значних помилок у розпізнаванні.

Аудіофайли, які мають низьку частоту дискретизації або зазнали стиснення з утратою якості, є складними для обробки. Такі файли

можуть містити спотворення звуку або втрату важливих мовних ознак, що робить їх важкими для точного аналізу.

Варіативність голосів: люди мають різні голоси, тембри, акценти та стилі мовлення. Ця варіативність додає складності для моделей, які повинні розпізнавати мовлення від різних спікерів. Моделі, які не були навчені на широкому діапазоні голосів, можуть демонструвати знижену точність при зміні спікера.

Акценти і діалекти: Особливо складними є акценти та регіональні діалекти, оскільки вони можуть мати інакшу вимову та фонетичні особливості, відмінні від стандартних мовних зразків, на яких навчалася модель.

Складність роботи з українськими даними в контексті розпізнавання мовлення

Однією з ключових проблем для систем розпізнавання мовлення (*ASR*) української мови є її складна граматика. Граматичні ознаки, як-от рід, число, відмінок, створюють велику кількість словоформ, ускладнюючи створення лексиконів. Наприклад, слово «друг» має варіанти «друга», «другові», «друзями» тощо, що значно збільшує кількість форм для обробки. Крім того, активне використання суфіксів і префіксів створює численні варіації слів, як-от «говорити», «пере-говорити», «роз-говорити». На прикладі чеської мови [6] показано, що для адекватного покриття усних і письмових текстів необхідно значно розширювати лексикон. Цей підхід може бути застосований і до української мови, враховуючи її багатство словоформ і відмінків.

Вільний порядок слів в українській мові також ускладнює моделювання, адже речення «Мама любить сина» і «Сина любить мама» мають однаковий зміст, але різну структуру. Це унеможливлює покладання на фіксовані мовні шаблони.

Також існує проблема багатомовності в Україні в контексті розпізнавання мовлення. Одна з унікальних особливостей мовної ситуації в Україні — це багатомовність, що створює додаткові виклики для систем автоматичного розпізнавання мовлення (*ASR*). Українці вдома спілкуються різними мовами, причому значна частина населення використовує й українську, й російську мови, а також змішані мовні форми. Це ускладнює завдання для моделей розпізнавання мовлення, які повинні бути здатні розпізнавати не лише українську мову, а й змішані мовні конструкції, що часто зустрічаються в розмовному стилі.

Згідно з даними опитування [7]:

- 44.7 % переважно говорять українською.
- 27,1 % спілкуються і російською, і українською однаково часто.
- 24,9 % українців переважно говорять російською.

Основні виклики багатомовності:

1. Мовний код-світчинг (*code-switching*). В Україні поширено явище, коли люди перемикаються між українською і російською мовами в рамках одного речення або навіть однієї фрази. Це створює виклики для *ASR*-систем, оскільки вони повинні бути здатними розпізнавати змішані мовні конструкції без втрати точності. Моделі, які розпізнають лише одну мову, можуть неадекватно реагувати на перемикання мов.

2. Російсько-український суржик. Значна частина населення України використовує так званий суржик — змішану форму мовлення, що поєднує елементи української та російської мов. *ASR*-системи, які були натреновані лише на стандартній українській або російській мові, можуть мати проблеми з розпізнаванням цієї змішаної мовної форми.

3. Різна частота використання мов у різних регіонах. У деяких регіонах України переважає російська мова, в інших — українська. Це означає, що системи розпізнавання мовлення мають бути адаптовані для різних регіональних мовних особливостей. Моделі, які недостатньо натреновані на таких регіональних особливостях, можуть демонструвати низьку точність у різних частинах країни.

4. Проблеми з акцентами та діалектами. Крім української, українці можуть спілкуватися регіональними діалектами, що додатково ускладнює задачу для систем *ASR*. Регіональні акценти можуть змінювати вимову слів, що робить

їх менш передбачуваними для системи розпізнавання.

Для ефективного розпізнавання мовлення в Україні, системи *ASR* повинні враховувати багатомовність та різні форми мовлення, які використовуються в країні. Це вимагає від моделей здатності працювати з кількома мовами одночасно, розуміти змішане мовлення (суржик) і враховувати регіональні мовні відмінності.

Час обробки моделей *whisper*

Моделі *Whisper* від *OpenAI* мають кілька версій, що відрізняються за розмірами та кількістю параметрів. Кожна з цих версій потребує різної обчислювальної потужності для обробки аудіо, що безпосередньо впливає на швидкість розпізнавання мовлення. На Рис. 4 можна побачити, як змінюється час обробки для різних версій *Whisper* на різних процесорах і графічних процесорах, що уможливлює оцінку ефективності кожної моделі залежно від наявних обчислювальних ресурсів [8].

Моделі *Whisper* доступні у версіях від *Tiny* до *Large-v3* (39 млн — 1,5 млрд параметрів). Збільшення параметрів покращує точність, але підвищує обчислювальні вимоги. Моделі *Tiny* і *Base* працюють швидше, але менш точні, тоді як *Large* і *Medium* забезпечують високу точність, проте не підходять для роботи в реальному часі на *CPU* через час обробки понад 15 секунд.

GPU, такі як *A*100, значно прискорюють обробку, але навіть на потужних графічних процесорах моделі *Large* і *Medium* не завжди досягають реального часу. Для мобільних пристроїв рекомендуються моделі *Tiny* або *Base*, які забезпечують швидку обробку (менш ніж 1 секунда на *CPU*), але знижують точність.

Загалом, Whisper — це потужний інструмент для розпізнавання мовлення, але час обробки даних вельми залежить від розміру моделі та доступних обчислювальних ресурсів. Великі моделі, такі як *Large* і *Medium*, забезпечують вищу точність, але потребують значної обчислювальної потужності, що робить їх менш придатними для роботи в реальному часі на звичайних *CPU* або мобільних пристроях.

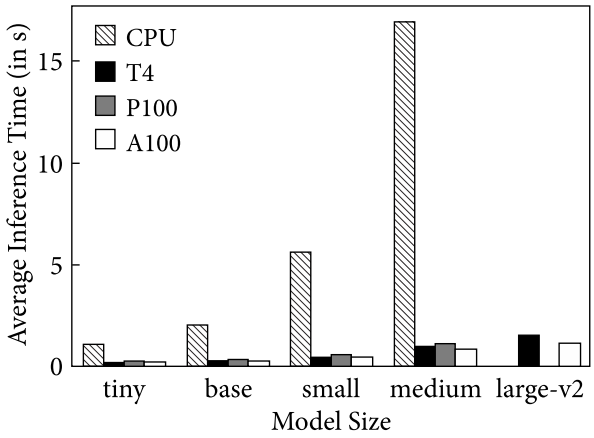

Рис. 4. Час роботи моделей *Whisper* залежно від типу обчислювального обладнання

Потенційний спосіб покращення швидкодії моделі *whisper*

Дистиляція є популярним методом у машинному навчанні, що дозволяє зменшити розмір великих моделей без значної втрати точності. У цьому процесі менша модель («учень») навчається на основі результатів більшої моделі («вчитель»), зберігаючи більшість її корисних властивостей. У контексті *ASR*, дистиляція оптимізує час обробки та ресурсомісткість, що є важливим для пристроїв з обмеженими ресурсами та задач реального часу [9].

Основні переваги дистиляції:

1. Зменшення розміру моделі: Наприклад, *Distil-Whisper* має на 51 % менше параметрів, що значно знижує вимоги до обчислювальних ресурсів.

2. Підвищення швидкості: *Distil-Whisper* стала у 5,8 разів швидшою, забезпечуючи ефективність для смартфонів та вбудованих систем.

3. Мінімальна втрата точності: Точність дистильованих моделей зменшується лише на 1—2 %, що робить їх придатними для більшості задач.

4. Енергоефективність: Завдяки меншому розміру, дистильовані моделі споживають менше енергії, що є ідеальним для мобільних пристроїв.

Дистиляція є важливим інструментом для задач реального часу, як-от голосові асистенти, транскрибування чи системи розпізнавання

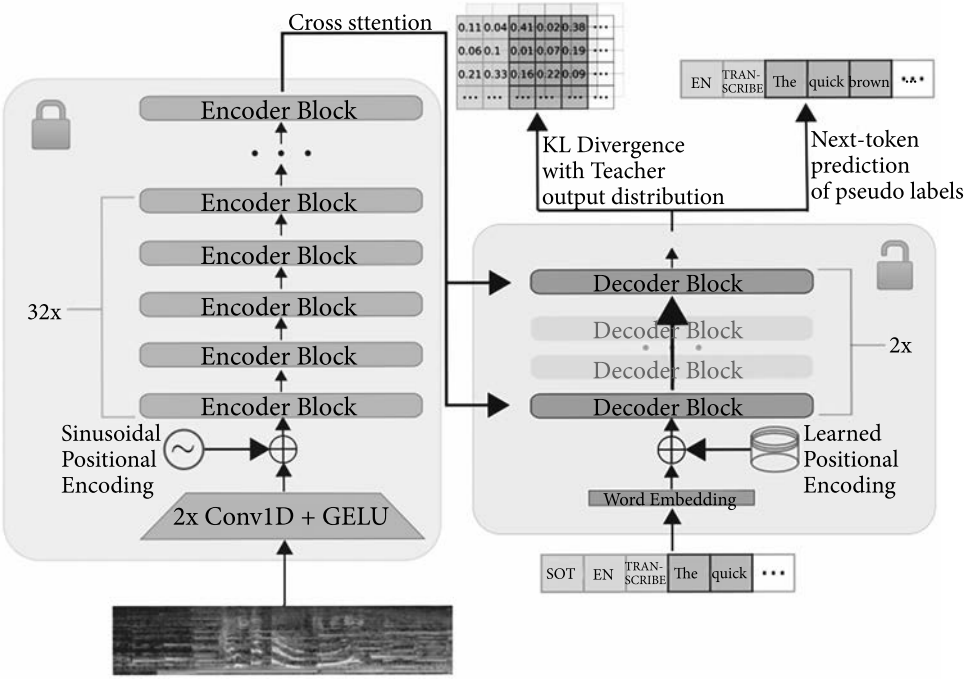

Рис. 5. Архітектура дистильованої *whisper*

мовлення, забезпечуючи баланс між швидкістю та точністю. Однак такі моделі можуть бути більш чутливими до шуму та обмеженими в багатомовних середовищах

Основна зміна учнівської моделі полягає у модифікації декодера, як можна побачити на рис. 5.

Особливості дистильованої архітектури моделі. *Whisper* має потужний енкодер, що складається з 32 блоків, який залишається незмінним під час дистиляції. Енкодер виконує основну роботу з витягування високоякісних ознак з аудіосигналу, що є критично важливим для точного розпізнавання мовлення. Це дає змогу зберегти високу точність моделі навіть після зменшення її розміру.

Декодер, який відповідає за перетворення вихідних ознак з енкодера на послідовність слів, був значно зменшений: кількість блоків зменшилася з 32 до 2. Це скорочення дає змогу значно прискорити роботу моделі, адже декодер тепер виконує менше обчислень, але при цьому ефективно використовує витягнуті ознаки, що забезпечує гарні результати.

Також особливістю є спосіб навчання через учительську та студентську модель. Навчання в такий спосіб використовує метод псевдо-розмітки, де вчительська модель допомагає навчати спрощений декодер, передаючи йому інформацію про прогнозовані наступні токени та розподіли вихідних даних. Це уможливлює максимальне наближення учнівської моделі до вчительської, навіть попри її зменшений розмір. Такий підхід дає змогу значно підвищити швидкість роботи моделі без суттєвої втрати точності, що робить дистильовані версії ефективними для використання в умовах обмежених ресурсів та реального часу. Проте потрібно не забувати про негативний вплив дистильованої моделі щодо можливості роботи з шумом.

Як видно з рис. 6 навіть при використанні дистиляції моделі, покращення *WER* значно залежить від кількості навчальних даних. При збільшенні обсягу даних з 435 годин до 21,770 годин, середній *WER* знижується з 16,4 % до 11,.4 %. Це показує, що чим більше даних доступно для навчання, тим краще модель справляється з розпізнаванням мовлення.

Але збільшення чутливості шуму не є найбільшою проблемою в рамках України. Адже більшість моделей після дистиляції втрачають багатомовність моделі, попри багатомовність вчительської моделі. Існують дослідження, пов'язані з дистиляцією знань *whisper-large-v2* для *whisper-small* [10], де в середньому було покращення точності для мов: Каталонська (*ca*), Чеська (*cs*), Галісійська (*gl*), Угорська (*hu*), Польська (*pl*), Тайська (*th*), Тамільська (*ta*), Українська (*uk*). В середньому дистильована модель показувала малу зміну позибки з 14.9 % до 16 %

Але якщо розглядати результати для української мови (Рис. 7), — то вони не були таки гарними. Адже результати для корпуса *FLEURS* (який не був тренувальною частиною корпуса) показали погіршення точності дистильованої моделі навіть порівняно з оригінальної whisper-small моделлю.

Size / h	Proportion / %	Avg. ID WER	Avg. OOD WER	Avg. WER
435	2	17.1	13.8	16.4
871	4	15.1	10.5	14.0
1,742	8	14.0	9.2	12.9
3,483	16	13.3	7.8	12.0
6,966	32	13.0	7.7	11.8
13,933	64	12.8	7.4	11.6
21,770	100	12.6	7.4	11.4

Рис. 6. Порівняння зменшення помилки дистильованої моделі в залежності від кількості годин для дистиляції

		FLEURS	CV-13
	#params	uk	uk
whisper-large-v2	**1.5B**	**8.1**	15.5
whisper-small	244M	20.5	32.3
whisper-small+FT	**244M**	**25.9**	**23.4**
whisper-small+LoRA-FT	379M	31.8	26.4
whisper-small+CLSR-FT	**369M**	**25.8**	**23.4**
DistilWhisper	369M	**24.9**	**23.1**

Рис. 7. Результати дистиляції моделі для української мови

Аналіз відкритих корпусів для української мови

Розглянемо перелік відкритих корпусів з українською транскрипцією (в тому числі з автоматичною).

Можна зробити висновок, що даних достатньо багато. Навіть без *Espreso TV subset* це майже 2900 годин транскрибованого аудіо. Але більшість із цих даних є автоматично транскрибованими, тож їхня якість не рівнозначна людській траскрибації.

Дослідження "*Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation*" [11] показує, як можна покращити розпізнавання мовлення з невеликими даними за допомогою самонавчання (*SSL*). У цьому методі модель вчиться на своїх власних транскрипціях.

Початковий корпус складав 24 хвилини даних і на них навчалась перша версія моделі. І згодом ця модель автоматично транскрибувала ще 168 хвилин даних. І фінальна модель почалась на 192 хвилині даних. Це покращило точність даних на 9,7 %. Але коли повторили цей експеримент уже з 96 хвилинами початкового корпусу — приріст складав вже 1,7 %.

З цього можемо зробити припущення, що після збільшення корпусу за допомогою автоматично транскрибованого датасету перестає давати значимий приріст до точності моделі. Але необхідні подальші дослідження для української мови

Таблиця 2. **Перелік відкритих корпусів з українською транскрипцією**

Корпус	Годин
Compiled dataset from different open sources + Companies + Community	1200
Voice of America	398
google/fleurs	12,2
YODAS	1147
M-AILABS	87
uk-pods	51
Espreso TV subset	23,0

Висновки

У рамках мого дослідження я дійшов висновку, що найкращою архітектурою для подальших досліджень у сфері автоматичного розпізна-

вання мовлення є трансформери. Серед них особливо перспективною є модель *Whisper* завдяки своїй стабільності та високим показникам точності, навіть у сценаріях *zero-shot*. *Whisper* показує видатні результати при роботі з мовами та умовами, для яких не було попереднього навчання.

Однією з ключових переваг цієї моделі є її навчання на величезних масивах даних, де значна частина складається з неанотованих даних. Це дозволяє *Whisper* бути стабільною до шуму й ефективно працювати у складних акустичних умовах. Попри розмір моделі, архітектура *Whisper* дозволяє проводити її дистиляцію з мінімальними втратами точності, що робить її придатною для обчислювально обмежених середовищ.

Досвід використання *Whisper* показав, що якість та кількість даних є одним із найважливіших факторів для покращення точності моделі. Успішне поєднання анотованих і неанотованих даних дозволяє моделі навчатися більш універсально, підвищуючи її стійкість до шумів та неточностей.

Таким чином, *Whisper* демонструє разючі можливості для подальших досліджень і розвитку систем автоматичного розпізнавання мовлення, особливо для української мови.

ЛІТЕРАТУРА

1. Jurafsky D., Martin J. Speech and Language Processing. 7 Jan. 2023. URL: https://web.stanford.edu/~jurafsky/slp3/A.pdf (дата звернення 01.08.2024).
2. Gales M., Steve Yo. "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in Signal Processing, 2007, vol. 1, no. 3, pp. 195–304. URL: https://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf (дата звернення 04.08.2024).
3. Jurafsky D., Martin J. Speech and Language Processing Automatic Speech Recognition and Text-To-Speech. URL: https://web.stanford.edu/~jurafsky/slp3/16.pdf (дата звернення 20.08.2024).
4. Vaswani A., et al. "Attention Is All You Need". ArXiv.org, 12 June 2017. URL: https://arxiv.org/abs/1706.03762 (дата звернення 20.08.2024).
5. Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. In International conference on machine learning). PMLR, 2023. pp. 28492–28518.
6. Nouza J., Zdansky J., Cerva P., Silovsky, J. Challenges in speech processing of Slavic languages (case studies in speech recognition of Czech and Slovak). Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23–27, 2009, Revised Selected Papers, pp. 225–241.
7. 24 Канал. "Якою мовою українці спілкуються вдома: опитування." 24 Канал, 17 Aug. 2021. URL: 24tv.ua/yakoyu-movoyu-ukrayintsi-spilkuyutsya-vdoma-opituvannya-ukrayina-novini_n1715078 (дата звернення 10.06.2024).
8. Shubham K. "Whisper Deployment Decisions: Part I — Evaluating Latency, Costs, and Performance Metrics." Medium, ML6team, 21 July 2023. URL: blog.ml6.eu/whisper-deployment-decisions-part-i-evaluating-latency-costs-and-performance-metrics-d07f6edc9ec0 (дата звернення 12.09.2024).
9. Gandhi S., von Platen P., Rush A. M. (2023). Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. arXiv preprint arXiv:2311.00430. 2023. URL: https://arxiv.org/abs/2311.00430 (дата звернення 01.09.2024).
10. Ferraz T. P., Boito M. Z., Brun C., Nikoulina V. Multilingual Distilwhisper: Efficient Distillation of Multi-Task Speech Models Via Language-Specific Experts. In ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024, pp. 10716-10720. DOI: 10.1109/ICASSP48485.2024.10447520
11. Bartelds M., San N., McDonnell B., Jurafsky D., Wieling M. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. ArXiv.org, 2023. URL: https://arxiv.org/abs/2305.10951 (дата звернення 26.08.2024).

Надійшла 13.09.2024

REFERENCES

1. Jurafsky, D., Martin, J. (2003) Speech and Language Processing. [online]. Available at: https://web.stanford.edu/~jurafsky/slp3/A.pdf [Accessed 1 Aug. 2024].
2. Gales, M., and Steve, Yo. (2007). "The Application of Hidden Markov Models in Speech Recognition." Foundations and Trends in Signal Processing, vol. 1, no. 3, pp. 195–304. [online]. Available at: https://mi.eng.cam.ac.uk/~mjfg/mjfg_NOW.pdf [Accessed 4 Aug. 2024].

3. Jurafsky, D., Martin, J. Speech and Language Processing Automatic Speech Recognition and Text-To-Speech. [online]. Available at: https://web.stanford.edu/~jurafsky/slp3/16.pdf [Accessed 20 Aug. 2024].
4. Vaswani, A., et al. "Attention Is All You Need". ArXiv.org, 12 June 2017, [online] Available at: https://arxiv.org/abs/1706.03762 [Accessed 20 Aug. 2024].
5. Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning). PMLR, pp. 28492–28518.
6. Nouza, J., Zdansky, J., Cerva, P., Silovsky, J. (2010). Challenges in speech processing of Slavic languages (case studies in speech recognition of Czech and Slovak). Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23–27, 2009, Revised Selected Papers, pp. 225–241.
7. 24 Канал. "Якою мовою українці спілкуються вдома: опитування." 24 Канал, 17 Aug. 2021, [online]. Available at: 24tv.ua/yakoyu-movoyu-ukrayintsi-spilkuyutsya-vdoma-opituvannya-ukrayina-novini_n1715078 [Accessed 10 Jun. 2024].
8. Shubham, K. "Whisper Deployment Decisions: Part I — Evaluating Latency, Costs, and Performance Metrics." Medium, ML6team, 21 July 2023. [online]. Available at: https://blog.ml6.eu/whisper-deployment-decisions-part-i-evaluating-latency-costs-and-performance-metrics-d07f6edc9ec0 [Accessed 12 Sept. 2024].
9. Gandhi, S., von Platen, P., Rush, A. M. (2023). Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. arXiv preprint arXiv:2311.00430. [online], Available at: https://arxiv.org/abs/2311.00430 [Accessed 1 Sept. 2024].
10. Ferraz, T. P., Boito, M. Z., Brun, C., Nikoulina, V. (2024). Multilingual Distilwhisper: Efficient Distillation of Multi-Task Speech Models Via Language-Specific Experts. In ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 10716–10720. DOI: 10.1109/ICASSP48485.2024.10447520
11. Bartelds, M., San, N., McDonnell, B., Jurafsky, D., Wieling, M. (2023). "Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation." ArXiv.org, 2023. [online]. Available at: https://arxiv.org/abs/2305.10951> [Accessed 26 Aug. 2024].

Надійшла 13.09.2024

Ye.R. Mrozek, PhD Student,
Department of Speech Recognition and Synthesis, International Research and Training Center for Information Technologies and Systems NAS and MES of Ukraine,
40, Akademika Glushkova Avenue, Kyiv, Ukraine, 03187,
ORCID: https://orcid.org/0009-0008-4989-5016,
zekamrozek@gmail.com

ANALYSIS OF MODERN APPROACHES
TO SPEECH RECOGNITION TASKS

Introduction. The necessity for modern approaches to solving speech recognition tasks arises from the rapid development of artificial intelligence and the need to improve the accuracy and speed of human-computer interaction in various areas, such as voice assistants, translation, and automation. This direction is becoming increasingly relevant due to the growing volume of generated audio data and the need for real-time processing, particularly in Ukrainian contexts where multiple languages and dialects coexist. Currently, several approaches to speech recognition, analysis, and transcription exist, including methods based on neural networks, speaker diarization techniques, noise removal, and data structuring. However, the challenge of creating a universal solution that meets the needs of multilingual environments and effectively handles unstructured audio data remains relevant.

Objective. To review existing tools and algorithms for solving speech recognition tasks, particularly for Ukrainian.

Methods. Speech recognition, deep learning, transformers.

Results. Theoretical foundations of approaches and models for speech recognition were considered for building a knowledge base for a multilingual spoken dialogue system. Effective examples of improving transcription accuracy for languages with limited data were also explored, along with potential steps to enhance system speed. Potential datasets for model training were discussed.

Conclusion. A structured review of modern methods for processing and analysing multilingual audio files was provided, outlining their advantages, disadvantages, and unresolved issues.

Keywords: *speech recognition, neural networks, machine learning.*