КОМПЮТЕРНІ ТЕХНОЛОГІЇ

УДК 004.056

МЕТОДИ ПІДВИЩЕННЯ ЯКОСТІ ПЕРЕТВОРЕННЯ МОВИ НА ТЕКСТ В СИСТЕМАХ БІОМЕТРИЧНОЇ АУТЕНТИФІКАЦІЇ

DOI 10.36994 / 2788-5518-2023-01-05-13

***Корчинський В.В.**, д.т.н.,проф. Державний університет інтелектуальних технологій і зв'язку, Одеса,Україна. vkadkorchin@ukr.net*

***Стайкуца С. В.,** к.ф.н.,доц. Державний університет інтелектуальних технологій і зв'язку, Одеса,Україна. s_t_a@ukr.net*

Виноградов І.В** Державний університет інтелектуальних технологій і зв'язку, Одеса, Україна**, ipvinner@gmail.com

***Швець О. В.** Державний університет інтелектуальних технологій і зв'язку, Одеса, Україна. ovshvets@ukr.net*

***Бєлова Ю. В.** Державний університет інтелектуальних технологій і зв'язку, Одеса, Україна. bilovaulia@gmail.com*

***Анотація:** У статті розглянуто методи та алгоритми перетворення мови на текст, сучасні відкриті та комерційні системи для створення систем, а також використання цих технологій у сфері кібербезпеки. Пропонується створити систему перетворення мови на текст високої якості. Проведено аналіз математичних алгоритмів, які використовуються для скорочення коефіцієнта помилок, що дає змогу створювати унікальні голосові відбитки та підвищити захист від підроблення. Описана структура сучасних систем перетворення мови на текст. Внаслідок зміни дата сетів, параметрів прихованих Марківських моделей, якісного словника фонем, використання мовних моделей існує можливість зменшити процент помилок при розпізнаванні мови, також використання системи для багатомовності типу "Суржик".*

Розглянуто математичні методи оцінки якості системи перетворення мови на текст(WER), а також різні методи розрахунку, що важливо для їхнього подальшого вдосконалення і оптимізації. Розглянуто структуру сучасних систем, а саме попередня обробка сигналу, витягування ознак, акустичне моделювання, моделювання мови, декодування, пост-обробка. Для кожного з этапів запропоновані вектори дослідження, які можуть зменшити коефіцієнт помилок системи в цілому.

Зменшення помилок розпізнавання мови та можливості підробити голос досягається за допомогою різних методів: глибокі нейронні мережі, приховані Марківські моделі, алгоритм Баума-Велша, N-gram моделі, Моделі з увагою, створення якісного словника фонем, датасету, філерів. Технологія перетворення мови в текст може бути використана в системах біометричної аутентифікації для виявлення та аналізу унікальних особливостей голосу користувача. Однак сучасні системи перетворення мови на текст для української, російської та "суржику" потребує вдосконалення, як акустичного, так і мовного блоку. Наукові роботи, які присвячені дослідженню та оптимізації цих систем для біометричної аутентифікації, не повною мірою висвітлюють ці питання. Це стало приводом для подальшого дослідження в цьому напрямку, тому метою цієї роботи є розробка системи розпізнавання мови з мінімальним коефіцієнтом помилок.

***Ключові слова:** перетворення мови на текст, мовні моделі, моделі прихованого марківського процесу, захист інформації, біометрична аутентифікація.*

METHODS OF IMPROVING THE QUALITY OF SPEECH-TO-TEXT CONVERSION IN BIOMETRIC AUTHENTICATION SYSTEMS

***Korchynskyi V.V.** , Dr. habil.,Prof. State University of Intellectual Technologies and Communication, Odessa, Ukraine, vkadkorchin@ukr.net*

***Staikuca S.V.,** Ph.D., Ass. Prof. State University of Intellectual Technologies and Communication, Odessa, Ukraine, s_t_a@ukr.net*

***Vynogradov I.V.** State University of Intellectual Technologies and Communications, Odessa, Ukraine, khaled@alfaiomi.com*

***Shvets O.V.** State University of Intellectual Technologies and Communication, Odessa, Ukraine, ovshvets@ukr.net*

Bielova I.V. *State University of Intellectual Technologies and Communications, Odessa, Ukraine, bilovaulia@gmail.com*

Abstract: *The article discusses the methods and algorithms of speech-to-text conversion, modern open and commercial systems for creating systems, as well as the use of these technologies in the field of cyber security. It is proposed to create a high-quality speech-to-text conversion system. An analysis of the mathematical algorithms used to reduce the error rate, which makes it possible to create unique voice prints and increase protection against forgery, has been carried out. The structure of modern speech-to-text conversion systems is described. By changing datasets, parameters of hidden Markov models, a high-quality dictionary of phonemes, and the use of language models, there is an opportunity to reduce the percentage of errors in language recognition, as well as the use of a system for multilingualism such as "surzhyk".*

The mathematical methods of assessing the quality of the system of speech to text (WER), as well as various methods of calculation, which is important for their further improvement and optimization, are considered. The structure of modern systems is considered, namely, signal pre-processing, feature extraction, acoustic modeling, speech modeling, decoding, post-processing. For each of the stages, study vectors have been proposed that can reduce the error rate of the system as a whole.

Reducing speech recognition errors and the ability to fake a voice is achieved using various methods: deep neural networks, hidden Markov models, Baum-Welch algorithm, N-gram models, models with attention, creation of a high-quality phonemes dictionary, dataset, and fillers. Speech-to-text conversion technology can be used in biometric authentication systems to detect and analyze the unique features of the user's voice. However, modern speech-to-text conversion systems for Ukrainian, Russian, and "surzhyk" need improvement in acoustic and language units. Scientific works, which are devoted to research and optimization of these systems for biometric authentication, do not fully cover these issues. This became the reason for further research in this direction, so this work aims to create a speech recognition system with a minimum error rate.

Keywords: *speech-to-text conversion, language models, hidden Markov process models, information protection, biometric authentication.*

Вступ

Технології перетворення мови на текст є важливим напрямком досліджень у сфері штучного інтелекту і мовознавства. Ці технології відіграють важливу роль у багатьох сферах[2], включаючи автоматичний переклад, синтез мовлення, підтримку осіб з обмеженими можливостями, аналіз тексту, автоматизації центрів обробки дзвінків та багато інших.

Одним з аргументів, які обґрунтовують необхідність дослідження цих технологій, є їх вплив на покращення комунікації між людьми з різними мовними навичками та культурними контекстами. Завдяки перетворенню мови на текст, можливо збільшити доступність інформації для широкого кола людей, незалежно від їхньої рідної мови чи здатності сприймати усну мову.

Дослідження[1,10] в цій області також допомагають вдосконалювати системи автоматичного перекладу, що має велике значення в глобалізованому світі, де спілкування між людьми різних національностей та мов є необхідністю. Застосування технологій перетворення мови на текст в автоматичному перекладі сприяє полегшенню комунікації, розвитку міжнародних бізнес-відносин та сприяє культурному обміну.

Технології перетворення мовлення на текст (Speech to Text) мають великий потенціал і можуть бути використані в Call центрах. Ось декілька способів, наприклад:

1) автоматичне розпізнавання мовлення: це може допомогти операторам Call центру ефективно працювати з великою кількістю дзвінків та забезпечувати точність та швидкість обробки клієнтських запитів;

2) розпізнавання інформації клієнта: ця технологія дозволяє автоматично виділяти ключову інформацію та використовувати її для подальшої обробки або маршрутизації дзвінків до відповідних відділів або операторів;

3) аналіз настрою та емоцій: ця технологія дозволяє оцінити задоволеність клієнта обслуговуванням та вчасно реагувати на негативні ситуації або проблеми, що виникають під час розмови;

4) автоматичне створення запису дзвінка: застосування технологій Speech to Text дозволяє автоматично транскрибувати дзвінки та створювати текстовий запис розмови.

Створення голосових асистентів передбачає застосовування технології перетворення мови на текст, що особливо актуально для певних завдань в галузі кібербезпеки декількома способами:

1) моніторинг спілкування: використовуючи технологію STT, компанії можуть відстежувати аудіо- та відеоконференції для виявлення потенційно шкідливої активності або обговорення

конфіденційної інформації. Це може бути особливо корисним для запобігання витоку інформації;

2) розпізнавання голосових команд: у сучасних системах, які використовують голосове управління, технології STT можуть аналізувати команди, щоб переконатися, що вони не містять потенційно шкідливих інструкцій;

3) біометрична аутентифікація: технології STT також можуть бути використані разом з технологією розпізнавання голосу для аутентифікації користувачів за їх голосом. Це може бути додатковим шаром захисту, на додаток до традиційних методів аутентифікації.

4) аналіз соціальних медіа: технології STT можуть бути використані для перетворення аудіо- та відео контенту в текст, який потім може бути проаналізований з метою виявлення загроз кібербезпеки. Наприклад, обговорення шкідливої активності або витоку конфіденційної інформації.

5) автоматизація реагування на інциденти: STT може бути використаний для автоматичного ведення записів в процесі реагування на інциденти, звільняючи спеціалістів кібербезпеки для важливіших завдань.

Таким чином, метою роботи є розробка системи розпізнавання мови з мінімальним коефіцієнтом помилок.

Виконання досліджень

Слід відзначити, що сучасні системи перетворення мови на текст (Speech to Text) мають певні недоліки, які можуть створювати також загрози інформаційної безпеки користувачів. Ось декілька з них:

1) недостатня точність: навіть найкращі системи перетворення мови на текст можуть містити неточності, особливо якщо розмова відбувається в шумному середовищі або якщо розмовники вимовляють слова невірно;

2) відсутність контексту: наприклад, якщо розмовники використовують слова з множинним значенням або вживають сленгові вирази, система може не зрозуміти, що саме має на увазі розмова;

3) питання безпеки: системи перетворення мови на текст можуть збирати та зберігати конфіденційну інформацію, яка може бути використана для шахрайства або ідентифікації користувачів;

4) обмеженість мов: багатомовні системи перетворення мови на текст можуть бути обмежені у підтримці різних мов та діалектів. Це може ускладнювати комунікацію з користувачами, які говорять іншими мовами або використовують місцеві діалекти. Особливо це актуально для України, де українська мова є державною мовою та є однією з основних мов, якою користуються українці. Російська мова також широко використовується. Крім того, в Україні проживає ряд етнічних груп, які зберігають свої рідні мови та мови-комуніканти, такі як румунська, польська, угорська, білоруська та інші;

5) помилки в інтерпретації мови.

Існуючі пропрієтарні та відкриті системи перетворення мови на текст мають певні недоліки та переваги. Пропрієтарні системи мають наступні особливості:

1) Google Speech-to-Text API – це сервіс від Google, який використовує передові алгоритми для перетворення мови на текст. Перевагами такого сервісу є висока точність розпізнавання, підтримка багатьох мов, Інтеграція з іншими сервісами Google. До недоліків можна віднести: вартість використання для великого обсягу даних; залежність від інтернет-з'єднання, питання приватності;

2) Nuance Communications Dragon – це один з найпопулярніших пропрієтарних продуктів для перетворення мови на текст. До переваг цього сервісу можна віднести високу точність розпізнавання, широкий вибір версій для різних професій, наявність програмного забезпечення для комп'ютерів та мобільних пристроїв. Недоліки сервісу – це висока вартість, обмежена підтримка мов, необхідність установки на локальний пристрій;

Можна відзначити наступні проєкти Open source, які характеризують системи перетворення мови на текст:

1) Mozilla DeepSpeech – це проект від Mozilla, який розробляється на базі технологій машинного навчання для перетворення мови на текст. Перевагами такого проєкту є відкритий код, можливість працювати на локальному пристрої, наявність пре-тренованих моделей для англійської мови. До недоліків можна віднести: нижча точність розпізнавання, порівняно з комерційними рішеннями,обмежена підтримка мов, можливі труднощі при встановленні та налаштуванні;

2) Kaldi – це відкритий проект для розробки систем розпізнавання мови на основі глибокого навчання. Перевагами проєкту є відкритий код, гнучкість та налаштованість, підтримка різних алгоритмів та методів машинного навчання;

3) CMU Sphinx – це один з найбільш відомих open source проектів для перетворення мови на текст, розроблений в Університеті Карнегі-Меллон. Перевагами цього проєкту є: відкритий код, що дозволяє налаштовувати систему під конкретні потреби, наявність версій для різних мов та доменів, можливість працювати на локальному пристрої без потреби в постійному інтернет-з'єднанні. До недоліків слід віднести: точність розпізнавання може бути нижчою в порівнянні з комерційними рішеннями, встановлення та налаштування може вимагати певних технічних знань, потреба в додаткових ресурсах для тренування моделей для нових мов або доменів[9].

Таким чином, одним із пунктів дослідження в роботі буде вибір відкритої системи для побудови власного алгоритму перетворення мови на текст. Як критерії ефективності виберемо математичні алгоритми, параметри яких планується удосконалити для отримання кращих результатів перетворення мови на текст, та знання мов програмування в автора.

Для завдання дослідження розглянемо наступні математичні алгоритми розпізнавання мови:

1) Hidden Markov Models (HMMs) може використовуватися для моделювання послідовності звуків у словах або фразах[5];

2) Gaussian Mixture Models (GMMs) використовується для представлення розподілу акустичних ознак в мові. Кожний компонент суміші може моделювати різні види звуків, такі як голосні або приголосні;

3) Viterbi Algorithm використовується для пошуку найбільш ймовірного шляху через HMM, що відповідає даному набору акустичних ознак[6];

4) Mel Frequency Cepstral Coefficients (MFCCs) забезпечує витягування ознак, який використовується для перетворення сигналу мови в набір ознак, які можна використовувати для тренування і тестування моделей розпізнавання мови;

5) N-gram Language Models використовуються для представлення ймовірностей послідовностей слів у мові. Наприклад, у моделі біграм, ймовірність кожного слова залежить від попереднього слова[5];

6) Baum-Welch Algorithm використовується для тренування параметрів HMM.

Таким чином, обґрунтованим є дослідити та за можливості оптимізувати один, або декілька алгоритмів для отримання кращих результатів перетворення мови на текст, а саме зниження WER, про який буде сказано далі.

Структура розпізнавання мови може бути розбита на декілька основних етапів:

1) попередня обробка: на цьому етапі, аудіосигнал, який містить мову, перетворюється в цифрову форму. Це включає в себе відлік сигналу, його фільтрацію та, можливо, зменшення шуму;

2) витягування ознак: одразу після попередньої обробки, аудіо сигнал аналізується з метою витягування корисних ознак, що характеризують мову. Тут використовуються техніки, такі як MFCC (Mel Frequency Cepstral Coefficients), що дозволяють перетворити сигнал в набір числових даних[2];

3) акустичне моделювання: на цьому етапі, використовуються акустичні моделі, зокрема Hidden Markov Models (HMMs) або Deep Neural Networks (DNNs), для визначення відповідності між ознаками аудіосигналу та фонемами, що є найменшими значущими звуковими одиницями мови[5];

4) моделювання мови: цей етап використовує моделі мови (зазвичай на основі n-грамів), які оцінюють ймовірність послідовності слів. Це допомагає вирішувати неоднозначності в розпізнаванні[5];

5) декодування: на цьому етапі, алгоритм вибирає найбільш ймовірну послідовність слів, враховуючи інформацію від акустичної моделі та моделі мови. Для цього часто використовується алгоритм Вітербі[8];

5) пост-обробка: після декодування, можливі деякі корекції або доповнення до тексту, такі як виправлення помилок, пунктуація тощо.

Система розпізнавання мови може бути розбита на два основних блоки: акустико-фонетичний та лінгвістичний.

Акустико-фонетичний блок – це частина системи, яка відповідає за обробку вхідного аудіо сигналу і його перетворення в послідовність фонем або інших звукових одиниць. Він складається з наступних етапів:

1) попередня обробка: на цьому етапі вхідний звуковий сигнал перетворюється в цифрову

форму, яка може бути оброблена комп'ютером. Це може включати відлік, фільтрацію та зменшення шуму;

2) витягування ознак: аудіо сигнал аналізується для витягування корисних ознак, що характеризують мову. Зазвичай використовується метод Mel Frequency Cepstral Coefficients (MFCC) для перетворення аудіо сигналу в набір числових даних;

3) акустичне моделювання: використовуються акустичні моделі, такі як Hidden Markov Models (HMMs) або Deep Neural Networks (DNNs), щоб визначити відповідність між ознаками аудіосигналу та фонемами[5].

Лінгвістичний блок – це частина системи, яка відповідає за моделювання структури мови та генерацію тексту. Він включає наступні етапи:

1) моделювання мови: на цьому етапі використовуються моделі мови (часто на основі n-грамів), щоб визначити ймовірність послідовності слів;

2) декодування: декодер використовує інформацію від акустичної моделі та моделі мови, щоб визначити найбільш ймовірну послідовність слів.

Розглянемо умова забезпечення якісного спектра сигналу для обробки та перетворення його в текст. Математичні моделі перетворення акустичного сигналу не є частиною цього дослідження, але оскільки вірогідно датасети будуть записані в умовах без шумів, треба дослідити можливості зменшення шуму.

Глибокі нейронні мережі (Deep Neural Networks, DNNs) можуть бути ефективними інструментами для зменшення шуму в акустичних сигналах, включаючи ті, які використовуються в системах автоматичного розпізнавання мови (ASR). Ці системи зазвичай можуть виправити або відфільтрувати шум в сигналі перед тим, як сигнал передається на більш високий рівень ASR для визначення фонем або слів[7].

Існує певні способи, за якими DNN можуть використовуватися для зменшення шуму:

1) Speech Enhancement: моделі DNN можуть використовуватися для визначення характеристик шуму в сигналі і зменшення його впливу. Наприклад, є така техніка як "Deep Noise Suppression", яка використовує глибокі нейронні мережі для визначення і видалення шуму з аудіосигналу;

2) Robust Acoustic Modeling: DNN також можуть бути натреновані для розпізнавання фонем в сигналах, які містять шум. Такі моделі, називаються стійкими акустичними моделями, намагаються визначити корисну інформацію в акустичному сигналі, незважаючи на наявність шуму;

3) Feature Compensation: корекція або компенсація ознак, які були спотворені шумом. Це включає використання DNN для виправлення або видалення шумових складових в ознаках акустичного сигналу.

Пошук, або створення якісного датасету для створення моделі розпізнавання мови є одним зі складних завдань цієї роботи. Нам знадобиться великий набір даних, який містить широкий спектр дикторів, акцентів, фраз та умов запису. Особливо це стосується суржику та інших змішаних варіацій мови. Щоб створити модель розпізнавання мови для суржику необхідно мати спеціалізований датасет, що містить аудіозаписи людей, які говорять цим діалектом. Такий датасет повинен включати велику кількість записів з різноманітними дикторами, контекстами, акцентами та швидкостями мовлення. Крім того, потрібні відповідні транскрипції цих записів, щоб модель могла вивчити відповідність між аудіосигналами та текстом. Певна робота вже була зроблена проєктом UA Silero, але на відміну від англійської чи російської мов, для української та суржику ще не існує якісних датасетов, які можна було б використовувати без доопрацювання. Також потрібно дослідити датасети, які вже є в open source, наприклад Common Voice від Mozilla або Voxforge і чи достатньо цього буде для завдання дослідження. Датасети відіграють важливу роль у створенні ефективних моделей розпізнавання мови з наступних причин та завдань:

1) навчання моделі: датасети використовуються для навчання моделей розпізнавання мови, які вивчають взаємозв'язки між аудіосигналами та відповідними транскрипціями тексту. Моделі використовують ці відповідності, щоб "вивчити" як перетворити аудіосигнали на текст;

2) різноманітність даних: ефективні моделі розпізнавання мови мають бути здатні розпізнавати мову в різних контекстах. Це означає, що вони повинні бути навчені на датасетах, які включають аудіо від різних дикторів з різними акцентами, голосами, швидкістю мовлення та іншими факторами;

3) тестування та валідація: датасети також використовуються для тестування та валідації моделей під час та після навчання. Це допомагає забезпечити, що модель працює правильно та здатна ефективно розпізнавати мову в реальних умовах.

Після отримання спектра, який простіше використовувати для системи розпізнавання мови потрібно буде знайти фонеми, філери та транскрипції. Для цього потрібно застосування словника фонем, словника філерів та словника транскрипцій:

1) словник фонем – це набір усіх фонем, що використовуються в певній мові. Фонема - це найменша одиниця звуку, яка може змінювати значення слова в мові. Одним із напрямків роботи буде спроба знайти або створити словник фонем для двох або більше мов, та спробувати використовувати його для побудови багатомовної системи перетворення мови на текст;

2) філери в системах Speech to Text (розпізнавання голосу) – це слова або звуки, які люди використовують у своєму повсякденному мовленні, але які не мають особливого значення для змісту висловлювання. Це слова, як "емм", "ну", "типу", "такий" і так далі. Використання філерів може підвищити якість розпізнавання на 10-15%;

3) словник транскрипцій – це ресурс, який використовується в процесі розпізнавання мови, щоб вказати, як слова мають бути вимовляються. Він зазвичай містить список слів разом з їхніми фонетичними транскрипціями.

Приховані марковські моделі (ПММ або HMM) вже тривалий час використовуються в системах розпізнавання мови[5] у розв'язанні ключових проблем статистичного розпізнавання, включаючи розпізнавання вимовленого слова, розпізнавання мови та розпізнавання диктора, а саме:

1) статистичний аналіз: ПММ дозволяють проводити статистичний аналіз розподілу ймовірностей для послідовностей слів, що допомагає у розпізнаванні мови.

2) моделювання часу: ПММ відображають часові залежності в даних. Це дозволяє їм моделювати послідовність звуків в словах, що є ключовим фактором в розпізнаванні мови.

3) розпізнавання слів: ПММ можуть бути навчені представляти конкретні слова або набори слів, що дозволяє системі розпізнавання мови ідентифікувати та класифікувати вимовлені слова.

4) розпізнавання диктора: ПММ можуть бути використані для ідентифікації особливостей голосу диктора, що поліпшує здатність системи розпізнавати та інтерпретувати мову різних людей.

Оптимізація HMM включає в себе ряд проблем, які вивчаються сучасними науковцями. Можна відзначити наступні проблеми, вирішення яких може удосконалити існуючі моделі перетворення мови на текст:

1) покращення алгоритму вивчення: в HMM використовується алгоритм Баума-Велша для вивчення параметрів моделі. Цей алгоритм є емпіричним, і сучасні науковці постійно шукають способи його покращення, щоб він був швидшим і точнішим;

2) оптимізація структури моделі: HMM може мати різні структури, залежно від контексту проблеми. Вибір найкращої структури моделі є однією з ключових проблем оптимізації HMM;

3) спрощення моделі: у багатьох ситуаціях, використання більш простих моделей може поліпшити швидкість та ефективність обчислень. Тому науковці досліджують способи спрощення моделей без значного зниження їхньої точності;

2) масштабування моделей: при використанні великих датасетів, може виникнути необхідність в масштабуванні моделей, що є іншою важливою проблемою оптимізації;

3) інтеграція з іншими моделями: інтеграція HMM з іншими статистичними або машинними моделями є ще однією областю дослідження;

3) універсальні моделі: розробка універсальних прихованих марковських моделей, які можуть бути ефективно використаними для перетворення;

4) мовні моделі є основними компонентами багатьох сучасних систем обробки природної мови.

За допомогою цих моделей вивчається структура тексту та генеруються ймовірні прогнози наступних слів або речень. Існує декілька типів мовних моделей:

1) N-gram моделі – це найпростіші мовні моделі, які використовують статистику з текстових даних для прогнозування ймовірності наступного слова на основі n-1 попередніх слів. Вони доволі прості у використанні, але мають обмеження, що стосуються віддалених залежностей в тексті та розріджених даних[5].

2) моделі прихованого марковського процесу (HMM) – ці моделі можуть бути використані для моделювання послідовностей слів. Такі моделі здатні обробляти віддалені залежності, але їхній алгоритм навчання може бути складним;

3) рекурентні нейронні мережі (RNN): RNN можуть обробляти довільну довжину послідовностей, що робить їх дуже корисними для мовних моделей. Однак, вони мають проблеми з віддаленими залежностями через проблему затухання градієнта;

3) моделі з увагою, як Transformer – засновані на механізмі "уваги", тобто ці моделі дозволяють моделювати залежності між словами, незалежно від їх відстані одне від одного в тексті[7];

4) моделі на основі BERT: BERT (Bidirectional Encoder Representations from Transformers) використовує Transformer архітектуру для створення потужних мовних моделей, які можуть моделювати контекст обох напрямків;

5) GPT (Generative Pre-training Transformer) – це серія моделей, розроблених OpenAI, що використовуються для розуміння і генерації природної мови.

Розглянемо можливості дослідження та оптимізації мовних моделей на основі *N*-gram. Ця модель є одна з найпростіших та найпоширеніших, в якій використовується машинне навчання для розпізнавання мови. При цьому відбувається поділ тексту на послідовність *N* слів та використання їх з метою вивчення ймовірностей мови.

Основні дослідницькі зусилля для оптимізації *N*-грам моделей включають:

1) згладжування: рроблема з N-грамами полягає в тому, що деякі послідовності слів можуть не з'явитися в тренувальному наборі даних, що призводить до нульової ймовірності. Згладжування допомагає вирішити цю проблему, шляхом призначення маленького значення ймовірності для таких послідовностей[5];

2) Backoff та interpolation: ці методи є способами управління ситуаціями, коли в тренувальних даних не вистачає інформації для оцінки ймовірності певного *N*-граму;

3) використання контексту: покращення способу, яким *N*-грами використовують контекст, може підвищити їхню точність. Замість того, щоб тільки використовувати безпосередній контекст N-1 попередніх слів, деякі дослідники дивляться на більш широкий контекст;

4) покращення обчислювальної ефективності: *N*-грам моделі можуть бути великими та складними, особливо для великих значень *N*. Дослідники шукають способи зменшення розміру моделі або покращення швидкості її обчислень[5];

5) використання глибокого навчання: нейронні мережі також можуть бути використані для побудови мовних *моделей* на основі *N*-грамів, здатних моделювати складніші залежності в даних;

6) комбінування з іншими методами: *N*-грам моделі часто використовуються в комбінації з іншими методами машинного навчання для створення більш потужних систем розпізнавання мови.

WER, або Word Error Rate, є загальноприйнятою мірою якості в задачах перетворення мови на текст, таких як автоматичне розпізнавання мови (ASR). WER вимірює кількість помилок, зроблених системою під час перетворення мовлення в текст, в порівнянні з референтним текстом.

WER розраховується на основі кількості вставок (*I*), видалень (*D*) і замін (*S*) необхідних для того, щоб перетворити гіпотезу (текст, згенерований системою ASR) в референсний текст. Він вимірюється за формулою:

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

де *N* – це кількість слів в референтному тексті.

Чим нижче WER, тим точніше система перетворення мови на текст. Проте WER також має свої обмеження, наприклад, він не враховує семантичний контекст або порядок слів, що можуть бути важливими в деяких застосуваннях.

Треба визначити кількість вставок, видалень та модифікацій, які необхідні для того, щоб розпізнана послідовність слів співпадала з еталонною послідовністю слів у текстовому корпусі. Цей процес обраховується на основі відстані Левенштейна.

Передові системи перетворення мови на тексти можуть досягати WER близько 5-10% на чистих, добре структурованих даних, наприклад, новинні відгуки або транскрипції подкастів. Однак в умовах реального світу, зі шумом у фоновому середовищі, акцентами, неформальною мовою, ймовірність помилки може збільшитись. Мета дослідження - створення системи перетворення української мови(опціонально російської та "суржику") яка буде мати якомога меньший WER.

Для оцінки треба використовувати перехресну валідацію. Кросвалідація, також відома як перехресна перевірка, це метод для оцінки якості моделі машинного навчання, що включає розділення даних на декілька підмножин та тестування моделі на одній підмножині після тренування на інших. В контексті автоматичного розпізнавання мови (ASR), кросвалідація може бути використана для оцінки того, наскільки добре модель може адаптувати своє навчання до нових даних. Наприклад, можна поділити набір даних для тренування на 10 підмножин (в методі,

що відомий як 10-кратна кросвалідація), потім навчати модель на 9 з них і тестувати її на 10-й. Це процес може бути повторений 10 разів, так що кожна підмножина використовується для тестування один раз.

Висновки

1) Хоча системи перетворення мови на текст вже досить ефективні, вони все ще мають ряд обмежень. Потрібно продовжувати дослідження і розробку, щоб поліпшити точність, ефективність і адаптивність цих систем.
2) Запропонований покроковий алгоритм створення системи speech to text, а також можливі шляхи покращення математичних, лінгвістичних та інженерних моделей для зменшення помилок перетворення мови на текст.
3) На першому етапі пропонується створити систему перетворення української мови з тих датасетів, які є у відкритому доступі. Збір даних - це не одноразовий процес. Для оптимальної роботи системи перетворення мови на текст датасет повинен бути постійно оновлюваний та покращуваний з урахуванням змін та покращення роботи системи
4) На другому етапі пропонується оптимізувати математичні алгоритми, які були оглядово розглянуті в цій статті для підвищення ймовірності визначення фонем.
5) На третьому етапі пропонується зі спеціалістами-філологами вдосконалити або створити якісні словники транскрипцій та словники фонем.
6) Отриманні результати будуть проаналізовані і використані для створення багатомовновної системи українська, суржик.

Література

1. A.Gaydhani, V.Doma, S.Kendre, L.Bhagwat Виявлення зневажливих та образливих слів у Twitter за допомогою машинного навчання: підхід на основі n-грам та статистичного показника (tf–idf). URL: https://arxiv.org/pdf/1809.08651.pdf
2. M.Suzuki, N.Itoh, T.Nagano, G.Kurata Удосконалення моделі мови n-грам з використанням тексту, згенерованого з моделі нейронної мови. URL: https://ieeexplore.ieee.org/document/8683481
3. R.Shadiev, WY.Hwang, NS.Chen, YM.Huang Огляд технології розпізнавання мовлення в текст для покращення навчання.URL: https://www.researchgate.net/publication/267811277_Review_of_Speech-to-Text_Recognition_Technology_for_Enhancing_Learning
4. DA.Jones, F.Wolf, E.Gibson, E.Williams, E.Fedorenko Вимірювання читабельності автоматичних розшифровок мовлення в текст. URL: https://www.researchgate.net/publication/221489176_Measuring_the_readability_of_automatic_speech-to-text_transcripts
5. BH.Juang, LR.Rabiner Приховані марковські моделі для розпізнавання мовлення. URL: https://www.jstor.org/stable/1268779
6. J.Picone - журнал IEEE Assp Безперервне розпізнавання мови з використанням прихованих марковських моделей. URL: https://ieeexplore.ieee.org/document/54527
7. Підходи до глибокого навчання в системі синтезу мовлення: систематичний огляд і перспектива останніх досліджень. URL: https://www.researchgate.net/publication/363982582_A_deep_learning_approaches_in_text-to-speech_system_a_systematic_review_and_recent_research_perspective
8. Y.Kumar, A.Koul, C.Singh - Мультимедійні засоби та програми, 2023 – Springer. URL: https://link.springer.com/article/10.1007/s11042-022-13943-4
9. Джерела та документи Sphinx. URL: https://www.sphinx-doc.org/en/master/
10. Lawrence Rabiner , Biing-Hwang Juang Основи розпізанавання мови. URL: https://www.academia.edu/4924307/Fundamental_of_Speech_Recognition_Lawrence_Rabiner_Biing_Hwang_Juang_

References

1. A Gaydhani, V Doma, S Kendre, L Bhagwat Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. URL: https://arxiv.org/pdf/1809.08651.pdf
2. M Suzuki, N Itoh, T Nagano, G Kurata Improvements to n-gram language model using text generated from neural language model.URL: https://ieeexplore.ieee.org/document/8683481
3. R Shadiev, WY Hwang, NS Chen, YM Huang Review of speech-to-text recognition technology for enhancing learning. URL: https://www.researchgate.net/publication/267811277_Review_of_Speech-to-Text_Recognition_Technology_for_Enhancing_Learning
4. DA Jones, F Wolf, E Gibson, E Williams, E Fedorenko Measuring the readability of automatic speech-to-text transcripts. URL: https://www.researchgate.net/publication/221489176_Measuring_the_readability_of_automatic_speech-to-text_transcripts
5. BH Juang, LR Rabiner Hidden Markov models for speech recognition. URL: https://www.jstor.org/stable/1268779
6. J Picone - IEEE Assp magazine Continuous speech recognition using hidden Markov models. URL: https://ieeexplore.ieee.org/document/54527

7. A deep learning approaches in text-to-speech system: A systematic review and recent research perspective. URL: https://www.researchgate.net/publication/363982582_A_deep_learning_approaches_in_text-to-speech_system_a_systematic_review_and_recent_research_perspective
8. Y Kumar, A Koul, C Singh - Multimedia Tools and Applications, 2023 – Springer. URL: https://link.springer.com/article/10.1007/s11042-022-13943-4
9. Sphinx sources and docs. URL: https://www.sphinx-doc.org/en/master/
10. Lawrence Rabiner,Biing-Hwang Juang Fundamentals of Speech Recognition. URL: https://www.academia.edu/4924307/Fundamental_of_Speech_Recognition_Lawrence_Rabiner_Biing_Hwang_Juang_