

РОЗПІЗНАВАННЯ ГОЛОСОВОГО ПОВІДОМЛЕННЯ У МЕСЕНДЖЕРІ

Стаття присвячена розгляду питання розпізнавання мовленнєвого сигналу. Описано етапи та елементи процесу обробки та розпізнавання природної мови з аудіосигналу. Наведено сучасні технології підтримки автоматичного розпізнавання мовлення та проблеми вибору серед них. Розглянуто аналіз останніх досліджень і публікацій щодо обробки голосових даних. Запропоновано рішення у вигляді чат-боту для автоматичного перетворення голосових повідомлень у текстові.

Ключові слова: розпізнавання мови, звукове повідомлення, ASR, NodeJS, месенджер, чат-бот.

M.P. Riadchenko, O.E. Piatykop. Recognizing speech in voice messages. The level of development of information technology makes it possible to use speech recognition technologies in a wide range of human life and activities. It is very convenient to use the voice interface: voice search for the necessary documents, dialing a phone number, managing IOT devices, voice navigation, simple text dictation. Since the natural language interface provides an additional convenience for a person when typing, sending voice messages has become common among users. In this case, voice messages are audio files. But it is not always available and convenient for the recipient to listen to such messages. This problem can be solved with the help of an automatic speech recognition system (ASR). The article describes the stages and elements of the process of processing and recognition of natural language by audio signal. Modern technologies of automatic speech recognition and problems with choosing among them are indicated. Modern automatic speech recognition (ASR) systems understand fully spontaneous speech that is natural, not memorized, contains signs of stuttering or even minor errors. At the same time, they are still too expensive to develop from scratch. So companies are faced with a choice between using the cloud API for ASR developed by the tech giants and using open source solutions. The analysis of the latest research and publications on the processing of voice data is considered. A software solution for automatic conversion of voice messages into text is proposed. The interface to the voice signal delivery system is proposed to be made as a chat bot in the messenger. The article presents the main components of the system, the algorithm of the chat bot, modern technologies for the development, implementation and configuration of the chat bot in the messenger.

Key words: speech recognition, audio, anti-noise improvement, ASR, NodeJS, messengers, chat-bot app.

Постановка проблеми. Рівень розвитку інформаційних технологій дозволяє використовувати технології розпізнавання мови у широкому колі життя та діяльності людини. Дуже зручно використовувати голосовий інтерфейс: голосовий пошук потрібних документів, набір номера телефону, управління ІОТ-пристроями, голосову навігацію, просте диктування тексту. Оскільки інтерфейс природною мовою забезпечує для людини додаткову зручність при наборі тексту, то серед користувачів поширеним стало надсилання голосових повідомлень. В цьому випадку голосові повідомлення представляють собою аудіо-файли. Але не завжди можливо та зручно прослуховувати такі повідомлення. А для адміністраторів каналів, лінгвістів, логопедів, викладачів та людей, що отримують інформацію з голосових повідомлень, прослуховування великої кількості яких по одному може обернутися надзвичайно неефективним надмірно трудомістким процесом.

¹ магістр, ДВНЗ «Приазовський державний технічний університет», м. Дніпро

² канд. техн. наук, доцент, ДВНЗ «Приазовський державний технічний університет», м. Дніпро, ORCID: 0000-0002-7731-3051, pee_pstu@ukr.net

Одним із способів вирішення цієї задачі є система підтримки віртуального діалогу завдяки чат-боту у популярних месенджерах: Telegram, Facebook Messenger, Slack. Зручний інтерфейс чат-боту допоможе користувачеві отримати з звукового файлу текстове повідомлення на екран. Але ядром програмної реалізації розпізнавання голосового повідомлення у месенджері є системи автоматичного розпізнавання мовлення (ASR). Загальний процес та елементи розпізнавання мовного сигналу показано на рисунку 1 [1, 2].

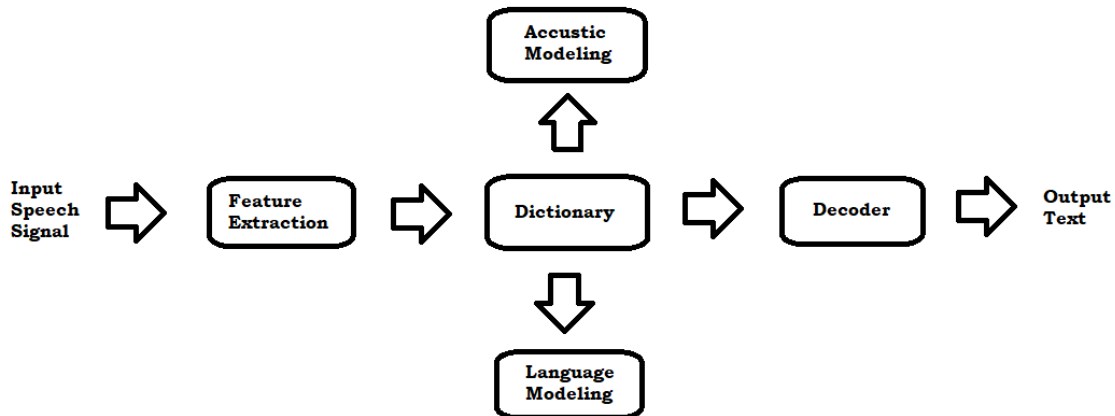


Рис. 1 – Кроки та елементи процесу розпізнавання мовленнєвого сигналу

На рисунку 1 наведено наступні елементи:

- Input Speech Signal (Аудіовхід) – мовленнєвий сигнал приймається в якості вхідного сигналу в систему за допомогою мікрофона, будь-якого подібного пристрою або з файлової системи;
- Feature Extraction (Виділення ознак) – цей блок виконує запис різних зразків мови. Оцифровка сигналу відбувається в цьому блоці, де виконується вибірка та квантування. Усунення шуму в мовленнєвому сигналі здійснюється шляхом квантування;
- Acoustic Modeling (Акустичне моделювання) – акустичне моделювання забезпечує статистичну модель мовленнєвого сигналу. Акустична модель призначає ймовірності фонемам, словами чи реченнями;
- Language Modeling (Мовне моделювання) – порівняння елементів з акустичної моделі з набором слів, присутніх у словнику, для отримання необхідної послідовності слів;
- Decoder (декодер) – використовується для розпізнавання точного слова.

Голосові повідомлення представляють спонтанну мову, яка є природною, не завченою, містить ознаки запинки чи навіть незначні помилки. Також при створенні голосових повідомлень може бути навколишній шум, який може впливати на точність результатів розпізнавання звукового повідомлення. Ще однією з актуальних проблем автоматичного розпізнавання мови є здатність обробляти голос користувача з різним набором акцентів. Тому дослідження в цьому напрямку є актуальними.

На сьогодні існує вже багато технологій створення інтелектуальних систем подібного типу - Google Cloud Speech API, Wit.ai, IBM Watson, Microsoft Azure Speech Services API та інші. Але комерційні системи пропонують обмежений доступ до деталізованих вихідних даних своїх моделей, включаючи матриці уваги (attention matrices), ймовірностей окремих слів та символів, вихідних даних проміжних рівнів, поряд з обмеженнями, пов'язаними з інтегруванням їх у інше програмне забезпечення. Тому у відповідь на ці обмеження з'являється все більше фреймворків ASR з відкритим вихідним кодом. Однак зростаюча кількість таких систем ускладнює розуміння того, яка з них найкраще відповідає потребам проекту, забезпечує повний контроль над процесом, яку можна використовувати без спеціальних знань у галузі глибокого навчання.

Тому при виборі між ASR вирішальним моментом є знаходження правильного балансу між зазвичай більш високою якістю пропріетарних систем і гнучкістю наборів інструментів з відкритим вихідним кодом. Всі ці ASR довели свою ефективність в різних умовах. Базові системи ASR

розпізнають записи з окремих слів, такі як відповіді «так» або «ні» і вимовлені цифри. Однак більш складні системи ASR підтримують безперервну мову і дозволяють вводити прямі запити або відповіді, такі як запит напрямку руху або номер телефону конкретного контакту. Якщо ASR використовується в звичайних і добре вивчених умовах і не вимагає занадто багато додаткової інформації, готова до використання система є найбільш оптимальним рішенням. Навпаки, якщо ASR є ядром проекту, більш гнучкий інструментарій з відкритим кодом стає кращим варіантом.

Метою даної роботи є дослідження моделей розпізнавання природної мови та існуючих хмарних автоматичних систем розпізнавання мовлення для їх придатності для інтегрування у месенджери для перетворення аудіоповідомлень у текст.

Аналіз останніх досліджень і публікацій.

Стаття [3] описує впровадження та оцінку ефективності мовлення ізольованого слова, незалежного від мовця. Для системи розпізнавання використовується прихована модель Маркова (HMM). Модель суміші Гауса використовується для моделювання розподілу мовних особливостей для кожного стану HMM. Система реалізована в MATLAB 7.9. Система навчена за допомогою власно створених баз даних, що складається з 60 зразків мовлення вибраних слів, перевірена на незалежний режим. Для обраних точність розпізнавання трьох слів 92%, 92% і 88% відповідно досягається в шумному середовищі.

У статті [4] «Speech Recognition in High Noise Environment» адресовано проблему розпізнавання у шумному середовищі та пропонується метод її вирішення. Більшість існуючих продуктів розпізнавання мови вимагають, щоб мова витягувалася за умови невеликого шуму або взагалі без шуму, що спрямовано на отримання правильної мовної інформації. Однак таке розпізнавання мови не можна використовувати у звичному середовищі. Так, наприклад, точність розпізнавання голосу різко знижується у рухомих автомобілях чи поїздах. Іноді голос взагалі неможливо розпізнати. Таким чином, діапазон адаптивності існуючого голосового продукту стає невеликим або непридатним. Це дослідження спрямоване на поліпшення результату розпізнавання мови у середовищі з підвищеним шумом. Пропонується новий метод, який використовує поліпшення мови у поєднанні з моделлю відкидання ознак. Новий метод дозволяє ефективно усунути вплив шуму на систему розпізнавання, незважаючи на складне середовище з великою кількістю помехового шуму. У цьому випадку правильна голосова інформація у повідомленні швидко ідентифікується. Також підвищується швидкість розпізнавання у автомобілі [4].

Розпізнавання мовлення дітей також є складним завданням [5]. Одна з причин полягає у тому, що дитяча мова має високу основну частоту, порівнянну зі значеннями частоти формантів. Крім того, у міру зростання дітей їх мовний апарат також зазнає змін. Це створює труднощі при надійному витяганні стандартних короткочасних спектральних характеристик для розпізнавання мови. В останні роки з'явилися нові методи акустичного моделювання, які дозволяють повністю очистити необхідні ознаки з необробленого мовного сигналу. Провівши дослідження на збірнику «PF-STAR corpus», проєкт «Improving Children Speech recognition Through Feature Learning From Raw Speech Signal» показує, що розпізнавання мовлення дітей може бути значно покращено з допомогою наскрізних методів акустичного моделювання [5].

Емоції відіграють життєво важливу роль у визначенні психічного стану людини. Емоції – це спосіб висловити свій світогляд та психічний стан іншим. У статті «Speech Emotion Recognition» [6] було проведено аналіз мови для вилучення емоцій. Вектор ознак містить компоненти аудіосигналу, які відображають специфічні особливості мовця, такі як тон, висота, енергія, що важливо для підготовки моделі класифікатора до точного розпізнавання конкретної емоції. Для створення набору даних для дослідницької роботи з мовленнєвих сигналів було проведено вилучення кепстральних коефіцієнтів малої частоти (MFCC). Вектор ознак, витягнутий з навчального набору даних, потім надсилається до моделі класифікатора. Тестовий набір даних пройде процедуру вилучення, після чого класифікатор зупиниться на виборі, що стосується емоції, прихованої у тестовому аудіо. Для підготовки набору даних було використано чотири різні набори даних (SAVEE, RAVDESS, TESS, CREMA-d). Використовувалися LSTM, випадковий ліс (метод машинного навчання для класифікації, регресії та інших завдань, який працює за допомогою побудови численних дерев прийняття рішень), давши точність майже 70% у тестовому наборі даних [6].

Зовсім інший підхід до розпізнавання емоцій було використано у статті «Speech Emotion Recognition Using Hidden Markov Models» [7]. У цій роботі представлена вдала спроба до

розпізнавання емоцій з допомогою системи розпізнавання RAMSES. Підхід заснований на стандартній технології розпізнавання мови з використанням прихованих напівнеперервних марківських моделей. Розглядаються як вибір низькорівневих функцій, так і дизайн системи розпізнавання. Наведено результати розпізнавання емоцій, що залежать від мовця, з використанням бази «Spanish Corpus of Interface Emotional Speech Synthesis». Точність розпізнавання 7 різних емоцій, визначених у форматі MPEG, перевищує 80% при використанні найкращого поєднання низькорівневих функцій та структури НММ. Цей результат дуже схожий на результат, отриманий з допомогою тієї ж бази даних при суб'єктивній оцінці суддями-людьми.

Виклад основного матеріалу. Система розпізнавання голосового повідомлення у месенджері складається з трьох основних компонентів:

- 1) сервер для обробки та зберігання даних;
- 2) модуль розпізнавання мови;
- 3) зручний інтерфейс для взаємодії користувача Messenger (для прикладу Telegram) та блоку розпізнавання.

Перший компонент (сервер) перетворює аудіозапис у необхідний формат, зберігає дані та надсилає перетворений файл до наступного модуля – служба розпізнавання мови. На даному етапі голосове повідомлення перетворюється на текст. Останній компонент – це з'єднувальний елемент між користувальницьким інтерфейсом та створеною службою трансформації.

В якості системи розпізнавання обрано хмарний сервісу AssemblyAI [8]. Платформа AssemblyAI – сукупність обчислювальних ресурсів, яка надається через послуги для широкої громадськості у вигляді публічної хмарної пропозиції. Ресурси платформи складаються з фізичної апаратної інфраструктури – комп'ютерів, жорстких дисків, твердотільних накопичувачів і мереж, що містяться у глобально розподілених центрах обробки даних, де будь-який з компонентів розгортається на замовлення, використовуючи шаблони, схожі на доступні у проекті Open Compute Project. Обладнання надається клієнтам у вигляді віртуалізованих ресурсів, таких як віртуальні машини.

Технологію можна використовувати для створення голосових інтерфейсів і транскрибування телефонних дзвінків. Як служба перетворення мовлення у текст цей інструмент може створювати розшифровки мовлення різними аудіоформатами і мовами. Система сприймає усі основні формати аудіо й автоматично перетворює їх на голосовий звук без будь-якого перекодування. Штучний інтелект збірки може підтримувати транскрипцію у форматі SRT або VTT як підписи або субтитри до відео. Однією з проблем транскрипції є граматики, тому AssemblyAI пропонує автоматичну пунктуацію та реєстр. AssemblyAI транскрипцію двоканального запису, тому користувачі матимуть окремі стенограми для кожного каналу. Усі ці функції підтримуються протоколами шифрування безпеки та конфіденційності, які включають «hard-видалення» тексту транскрипції з бази даних AssemblyAI [8].

Для взаємодії користувача з сервісом розпізнавання мови за допомогою «BotFather» [9] було створено зручний інтерфейс у вигляді чат-боту на базі популярного кроссплатформенного месенджера Telegram.

Telegram-бот – спеціальний обліковий запис, який створюється без прив'язки до номера телефону для автоматичної обробки і надсилання повідомлень у розмові або груповому чаті. Спілкування між такими аккаунтами організовано з допомогою звичайного інтерфейсу HTTPS спрощеними методами Telegram API [10]. Основним призначенням облікового запису є надання інтерфейсу до розробленого сервісу, який працює на хмарному сервері. Спілкування з ботом відбувається з допомогою команд, що повинні починатися з символу косої межі «/» і не можуть бути довші за 32 символи. Синтаксис команд виглядає наступним чином: /команда [необов'язковий] [аргумент].

Алгоритм роботи такого боту складається з етапів:

- формування посилання з допомогою Axios [11] (бібліотека з відкритим вихідним кодом заснована на промісах ES6, що серед іншого дозволяє робити HTTP запити до сервера, надаючи методи GET, POST, DELETE) На стороні сервера використовується власний HTTP-модуль NodeJS [12], тоді як на стороні клієнта – XMLHttpRequests.

- отримання доріжки (використовуючи власний модуль NodeJS – fs для операцій вводу-виведення) з отриманого або пересланого звукового повідомлення. Якщо вона відсутня, бот

свідомо не повідомляє про помилку, адже це призводить до великої кількості службових повідомлень у групах, якими користуються потенційно тисячі користувачів-не ІТ-спеціалістів).

- перетворення у потрібний формат: ayaabuffer, що являє собою по суті зліпок пам'яті комп'ютера без змін, у формат OGG – мультимедіа-контейнер для зберігання потоків даних, таких як відео, аудіо і субтитри. Він забезпечує більш надійне транспортування файлу, контроль цілісності файлу, мінімізацію кількості переміщень по файлу під час відтворення є у декількох потоках.

Створеному телеграму-боту можна відправити або переслати голосове повідомлення та у відповідь отримати текст. Приклад показано на рисунку 2.

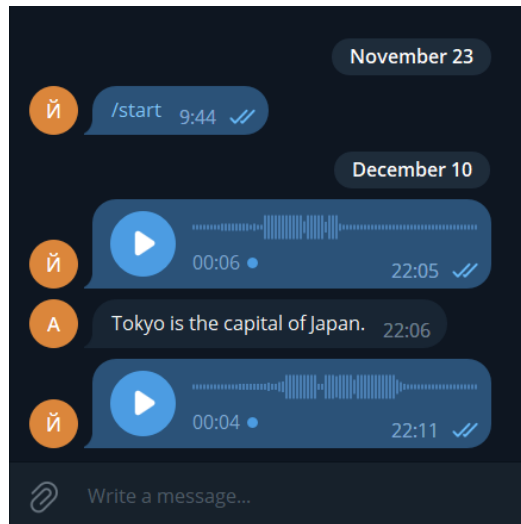


Рис. 2 – Режим взаємодії з ботом через діалог.

Після розробки було проведено оцінку роботи даної системи за такими параметрами, як точність розпізнавання і швидкість розпізнавання мовлення. За точність розпізнавання прийнято відношення кількості записів у навчальній базі до кількості правильно розпізнаних слів. За швидкість розпізнавання прийнято час між завершенням вимови слова та моментом видачі системою результату розпізнавання. Середня швидкість склала 3280,5 мс. Результати наведені у таблиці.

Таблиця

Результати оцінки роботи системи перетворення голосових повідомлень у текстові

Слово	Кількість варіантів	Коректно класифіковано	Точність (%)	Швидкість розпізнавання, ms
One	50	49	98%	3267
Two	50	47	94%	2919
Three	50	46	92%	3050
Four	50	45	90%	3172
Five	50	47	94%	3037
Six	50	47	94%	3366
Seven	50	50	100%	3641
Eight	50	49	98%	4146
Nine	50	48	96%	3291
Ten	50	47	94%	3318
Stop	50	48	96%	3057
Left	50	50	100%	3309
Right	50	47	94%	3074
Всього	650	619	95,2%	3280,5

Висновки

У роботі було виконано дослідження моделей розпізнавання природної мови та існуючих хмарних автоматичних систем розпізнавання мовлення для їх придатності ASR до потреб інтегрування у програмні продукти інтелектуальних систем подібного типу. Під час роботи проведено аналіз наукових досліджень та публікацій, сучасних технологій та програмного забезпечення, що присвячені розпізнаванню звукових повідомлень природної мови.

Запропоновано програмне рішення автоматичного перетворення голосових повідомлень у текстові на базі чат-боту для платформи Telegram. Оцінені результати точності та швидкості розпізнавання системи. Таким чином можна зробити висновок про доцільність використання запропонованої системи.

Перелік використаних джерел:

1. Добрушкін Г.О. Основні підходи до розпізнавання мовленнєвої інформації (частина 1) / Г.О. Добрушкін, В.Я. Данилов // Вісник Вінницького політехнічного інституту. – 2009. – № 47. – С. 50-64
2. Васильєва Н.Б. Проблеми створення систем розпізнавання мовлення для різних комп'ютерних платформ / Н.Б. Васильєва, Д.Я. Федорин // Штучний інтелект. – 2013. – Вип. № 4. – С. 158-167.
3. Chavan Rupali S. An Implementation of Text Dependent Speaker Independent Isolated Word Speech Recognition Using HMM Ms / Rupali S. Chavan, Dr. Ganesh S. Sable // Journal of Engineering Sciences & Research Technology. – 2013. – Vol. 2(9). – Pp. 2311-2318.
4. Tang C. Speech Recognition in High Noise Environment / C. Tang, M. Li // Ekoloji. – 2019. – Vol. 28(107). – Pp. 1561-1565.
5. Dave N. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition / N. Dave // International Journal For Advance Research in Engineering And Technology. – 2013. – Vol. 1, iss. VI. – Pp. 1-5.
6. Dubagunta S.P. Improving Children Speech Recognition through Feature Learning from Raw Speech Signal / S.P. Dubagunta, S.H. Kabil, Doss M. Magimai // 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2019. – Pp. 5736-5740. – Mode of access: <https://doi.org/10.1109/ICASSP.2019.8682826>.
7. Mittal R. Speech Emotion Recognition / R. Mittal, S. Vart // 2nd International Conference on Intelligent Technologies (CONIT). – 2022. – Pp. 1-6. – Mode of access: <https://doi.org/10.1109/CONIT55038.2022.9848265>.
8. AssemblyAI API Platform for Models [Electronic resource]. – Mode of access: <https://www.assemblyai.com>.
9. Telegram Bot Features [Electronic resource]. – Mode of access: <https://core.telegram.org/bots/features>.
10. Bot API Reference [Electronic resource]. – Mode of access: <https://tigrm.ru/docs/bots/api>.
11. Axios [Electronic resource]. – Mode of access: <https://axios-http.com>.
12. Node.js | About [Electronic resource]. – Mode of access: <https://nodejs.org/about>.

References:

1. Dobrushkin G.O., Danilov V.Ia. Osnovni pidkhodi do rozpiznavannia movlennivoi informatsii (chastina 1) [Basic approaches to recognizing speech information (part 1)]. *Visnik Vinnits'kogo politekhnichnogo institutu – Visnyk of Vinnytsia Polytechnical Institute*, 2009, № 47, pp. 50-64. (Ukr.)
2. Vasil'eva N.B., Fedorin D.Ia. Problemi stvorennia sistem rozpiznavannia movlennia dlia riznykh komp'uternykh platform [Problems of creating speech recognition systems for different computer platforms]. *Shtuchnyi intelekt – Artificial Intelligence*, 2013, vol. № 4, pp. 158-167. (Ukr.)
3. Chavan Rupali S., Sable Dr. Ganesh S. An Implementation of Text Dependent Speaker Independent Isolated Word Speech Recognition Using HMM Ms. *Journal of Engineering Sciences & Research Technology*, 2013, vol. 2(9), pp. 2311-2318.
4. Tang C., Li M. Speech Recognition in High Noise Environment. *Ekoloji*, 2019, vol. 28(107), pp. 1561-1565.

5. Dave N. Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition. *International Journal For Advance Research in Engineering And Technology*, 2013, vol. 1, iss. VI, pp. 1-5.
6. Dubagunta S.P., Kabil S.H., Magimai Doss M. Improving Children Speech Recognition through Feature Learning from Raw Speech Signal. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5736-5740. doi: **10.1109/ICASSP.2019.8682826**.
7. Mittal R. Speech Emotion Recognition / R. Mittal, S. Vart // 2nd International Conference on Intelligent Technologies (CONIT). – 2022. – Pp. 1-6. – Mode of access: doi: **10.1109/CONIT55038.2022.9848265**.
8. AssemblyAI API Platform for Models Available at: www.assemblyai.com (accessed 10 May 2022).
9. Telegram Bot Features [Electronic resource]. – Mode of access: <https://core.telegram.org/bots/features> (accessed 15 May 2022).
10. Bot API Reference [Electronic resource]. – Mode of access: <https://tigrm.ru/docs/bots/api> (accessed 25 April 2022).
11. Axios [Electronic resource]. – Mode of access: <https://axios-http.com> (accessed 20 May 2022).
12. Node.js | About [Electronic resource]. – Mode of access: <https://nodejs.org/about> (accessed 25 May 2022).

Рецензент: О.І. Проніна
канд. техн. наук, доцент, ДВНЗ «ПДТУ»

Стаття надійшла 30.09.2022