ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Харківський національний університет радіоелектроніки

Програмні засоби розпізнавання голосових повідомлень за допомогою машинного навчання

Виконав:
здобувач 4 року
навчання
групи КІУКІз-21-1
Дудник В.О.

Харків-2025

Мета та завдання дослідження

Мета:

- Розробка системи розпізнавання голосу для ідентифікації особи.

Завдання:

- Проаналізувати існуючі методи розпізнавання мовлення
- Вивчити доступні програмні інструменти
- Розробити математичну модель
- Реалізувати систему з використанням класифікаторів
- Провести тестування та оцінку

Актуальність теми

- Біометрична безпека - індивідуальні фізіологічні або поведінкові характеристики людини, які важко підробити. Голос, як елемент поведінкової біометрії, дає змогу ідентифікувати людину навіть дистанційно, що особливо актуально для банківських сервісів, «розумних» будинків, телемедицини, віддаленої роботи тощо.
- Сучасні методи ML/DL дозволяють досягти високої точності - завдяки машинному навчанню (ML) та глибокому навчанню (DL) моделі навчаються на тисячах годин мовлення та самостійно виявляють закономірності. Наприклад, нейромережі типу RNN, CNN або трансформери (як у Whisper від OpenAI) здатні розпізнавати голос із точністю понад 95% у складних умовах

3

Загальна теорія

- Використано Мел-частотні кепстральні коефіцієнти (MFCC) для вилучення ознак, також, вони дозволяють виділити важливі акустичні характеристики, що відрізняють різні звуки.
- Застосовані такі нейронні класифікатори: Багатошаровий персептрон (MLP) , Мережа радіальних базисних функцій (RBFN), Метод опорних векторів (SVM), Випадковий Ліс, Наївний Байєс, Гаусова модель суміші.
- Модель верифікації реалізовано на базі моделі гауссової суміші (GMM) - кожна точка даних має певну ймовірність належати до кожного з цих гауссових розподілів, що дозволяє GMM виконувати м'яку кластеризацію.

4

Переваги та недоліки існуючих програм для розпізнавання мовлення

- Amazon Transcribe – Переваги: не потрібно вручну обробляти аудіофайл, підтримується багато форматів аудіофайлів. Недоліки: він не розпізнає числові розряди як вимовлені; він перетворює їх на «один» або «два» замість 1, 2.
- Microsoft Bing Speech API – Переваги: дуже простий у використанні. Недоліки: Переклад може бути дивним, але зміст зрозумілий.
- Whisper – Переваги: програмне забезпечення з відкритим вихідним кодом і дуже вигідна ціна при використанні (0,006$ за хвилину). Недоліки: якщо у нас є повна транскрипція, то модель не може повністю транскрибувати її за один раз, оскільки вона розроблена для використання лише 30 секунд аудіофайлу.
- Otter.ai – Переваги: Можна повністю зосередитися на тих, з ким спілкуюся під час дзвінка, не маючи потреби постійно робити нотатки, Otter робитиме нотатки та записуватиме аудіостенограму. Недоліки: Іноді точність транскрипції падає через сильні акценти або фоновий шум.

5

Реалізація

- Видалено фонові шуми та нормалізовано рівень сигналу за допомогою Adobe Audition.
- Із сигналів були отримані MFCC, а також Delta та Delta-Delta коефіцієнти — як ключові ознаки для розпізнавання.
- Для ідентифікації мовця застосовано декілька алгоритмів машинного навчання: Multilayer Perceptron (MLP), Radial Basis Function Network (RBFN), Support Vector Machine (SVM), Random Forest.
- Реалізовано початкове розділення за статтю мовця для підвищення точності подальшої ідентифікації.
- Впроваджено модель GMM, яка оцінює ймовірність належності зразка до конкретного користувача.
- Система реалізована на Python із використанням бібліотек: libros, sklearn, numpy, matplotlib.

6

Реалізація (Python)

- Система розпізнавання голосу була реалізована на мові програмування Python із використанням таких бібліотек як librosa, scikit-learn, numpy та matplotlib. На першому етапі відбувався збір аудіозаписів у форматі .wav, які попередньо очищались від шуму, нормалізувались та обрізались до однакової довжини. Для кожного запису виділялись ознаки мовлення, зокрема MFCC, Delta та Delta-Delta коефіцієнти, що в сумі давало 39 параметрів на фрейм.
- Отримані ознаки масштабувались за допомогою StandardScaler або MinMaxScaler, після чого використовувались для навчання моделей класифікації. Спочатку проводилась класифікація статі мовця із застосуванням таких алгоритмів як MLP, SVM, Random Forest, KNN, Decision Tree та Naive Bayes. Далі, в межах виявленої статі, здійснювалась ідентифікація конкретного мовця на основі навченої моделі.
- На завершальному етапі було реалізовано верифікацію особи за допомогою Gaussian Mixture Model. Для кожного користувача створювалась окрема GMM-модель, яка оцінювала ймовірність того, що новий зразок належить саме цьому мовцю. Такий підхід дозволив поєднати класифікацію та верифікацію в єдину систему розпізнавання мовця з високою точністю

7

Таблиця 3.1 - Аналіз розпізнавання статі та порівняння між 9 класифікаторами в наборі даних

Алгоритми	З масштабуванням (точність) дані були нормалізовані		Без масштабування (точність) навчання без попередньої нормалізації ознак	
	Навчання	Тестування	Навчання	Тестування
1	2	3	4	5
Багатошаровий персептрон	0.65	0.64	0.65	0.64
Радіальна базисна функція	0.35	0.36	1.00	0.38

1	2	3	4	5
Дерево рішень	1.00	1.00	1.00	1.00
Випадковий ліс	1.00	1.00	1.00	1.00
Посилення градієнта	1.00	1.00	1.00	1.00
Наївний Байес	1.00	1.00	1.00	1.00
К-Найближчий сусід	0.39	0.52	0.81	0.69
Логістична регресія	0.35	0.36	0.82	0.62
Підтримуюча векторна машина	0.35	0.36	1.00	1.00

8

Таблиця 3.2 - Аналіз розпізнавання статі та порівняння між 7 класифікаторами в наборі даних Kaggle

Алгоритми	З масштабуванням (точність)		Без масштабування (точність)	
	Навчання	Тестування	Навчання	Тестування
Багатошаровий персептрон	0.63	0.63	0.92	0.91
Дерево рішень	0.52	0.52	0.80	0.80
Випадковий ліс	1.00	0.67	1.00	0.98
Посилення градієнта	1.00	0.67	1.00	0.98
Наївний Байєс	0.42	0.42	0.67	0.67
K-Найближчий сусід	0.60	0.60	0.99	0.96
Логістична регресія	0.58	0.58	0.89	0.89

9

Результати

- Найвищу точність (100%) як на навчанні, так і на тесті показали: Дерево рішень, Випадковий ліс, Посилення градієнта, Наївний Байєс. Це свідчить або про добре навчання, або про потенційне переобучення.
- Радіальна базисна функція - без масштабування: 100% точність на навчанні, 38% на тесті. Це означає сильне переобучення, модель запам'ятала приклади, але не навчилась узагальнювати.
- Підтримуюча векторна машина - з масштабуванням: (~35%) без масштабування: 100% точність на обох наборах. У даному випадку масштабування ознак шкодило SVM, і вихідні дані вже були у зручному для нього діапазоні.
- К-найближчий сусід- з масштабуванням:52% без масштабування: 69%. Дуже чутлива до відстаней, тому масштабування ознак важливе, але залежить від розподілу.

10

Висновки

Досягнута висока точність у визначенні статі мовця дозволяє оптимізувати ідентифікацію особи шляхом обмеження пошуку до відповідного підрозділу бази даних. Під час реалізації було використано декілька класифікаторів, серед яких деякі продемонстрували значну ефективність. Це відкриває можливості для застосування такої системи у масштабних сценаріях, включаючи корпоративні або національні рішення у сфері безпеки.

Подальше вдосконалення можливо шляхом інтеграції більш вдосконалих алгоритмів зіставлення зразків, масштабування системи для обробки великих обсягів даних, а також реалізації гнучкої логіки пошуку в базі даних у разі невизначеності під час класифікації статі.