УДК 004.934:004.032.26

ПОРІВНЯННЯ СИСТЕМ РОЗПІЗНАВАННЯ ГОЛОСУ ОСНОВАНИХ НА СТАТИСТИЧНІЙ МОДЕЛІ І ГЛИБОКІ НЕЙРОННІ МЕРЕЖІ

Мазепа А.С.
Науковий керівник – канд. техн. наук, доц. Єсілевський В.С.
Харківський національний університет радіоелектроніки, каф. ПМ,
м. Харків, Україна
тел. +38(050) 580-11-89, email: andrii.mazepa@nure.ua

This work provides an overview of different approaches to voice recognition as statistical models like Hidden Markov Models and deep neural networks like CNNs, RNNs, and LSTMs withthe challenges of noise in voice recognition, and the benefits of DNN-based models in handling noisy environments.

Традиційні статистичні моделі: приховані марківські моделі (HMM) і моделі n-грам, важко справляються з високим рівнем шуму.

Розглянемо ці підходи і порівняємо їх ефективність у завдання розпізнавання голосу текст. Також на практиці порівняємо одні з найпопулярніших програмних пакетів (моделі n-грам і DNN).

HMM представляють мову як послідовність прихованих станів [1]:

$$P(O \mid S) = \sum_{p,\, q} P(O \mid q) P(q \mid S),$$

де O – аудіосигнал, що спостерігається, S – послідовність прихованих станів, q – конкретний стан у послідовності, $P(O \mid q)$ – ймовірність аудіосигналу при заданому стані q, а $P(q \mid S)$ є ймовірність стану q при заданій послідовності S.

Моделі n-грам оцінюють ймовірність наступного слова у послідовності з урахуванням попередніх слів [2]:

$$P(O \mid W) = \prod_{t} P(w_t \mid w_{t-n+1}, ..., w_{t-1}),$$

де O – звуковий сигнал, що спостерігається, W – послідовність попередніх слів, w_t – поточне слово, а $P(w_t \mid w_{t-n+1}, ..., w_{t-1})$ – ймовірність поточного слова з урахуванням попередних $n-1$ слів.

Традиційні статистичні моделі, такі як HMM та моделі n-грам, мають обмеження в роботі з шумним середовищем та акцентами.

Системи розпізнавання мовлення на основі DNN використовують різні архітектури нейронних мереж для отримання ознак з аудіосигналу та моделювання часових залежностей у мовленні. Один з підходів полягає у використанні CNN для вилучення акустичних характеристик з необробленого аудіосигналу, потім передати RNN, наприклад LSTM. Потім вихідні дані LSTM передаються через повністю підключену нейронну мережу виконання завдання класифікації [3], [4].

CNN (Convolutional Neural Network), тип нейронної мережі, який за-

звичай використовується для завдань обробки зображень і сигналів.

RNN – рекурентна нейронна мережа.

LSTM (Long Short-Term Memory), тип RNN, призначений для вирішення проблеми градієнтів, що зникають, у традиційних RNN.

Порівняння статистичних моделей та DNN:

Результати тестових прогонів HTK та Kaldi на тестових наборах Verbmobil 1 corpus та Wall Street Journal 1 corpus показані в таблиці 1 [5].

У порівнянні з іншими розпізнавальниками визначну продуктивність Kaldi можна розглядати як революцію в технології розпізнавання мовлення з відкритим кодом.

HTK – це складний набір інструментів для роботи. Конвеєр навчання забирає багато часу і схильний до помилок.

Таблиця 1 – Частота помилок у словах на тестовому наборі VM1 та тестовому наборі WSJ1 за листопад 1993 р.

recognizer	VM1	WSJ1
HDecode v3.4.1	22.9	19.8
Julius v4.3	27.2	23.1
Kaldi	12.7	6.5

HDecode, Julius та Kaldi – це три набори інструментів з відкритим вихідним кодом, які використовуються для створення та навчання систем розпізнавання мовлення. Kaldi вважається найпопулярнішим і зручним для користувача набором інструментів із набором попередньо навчених моделей та прикладів сценаріїв.

Список використаних джерел:

1. Jurafsky, D., & Martin J.H. (2020). *Speech and Language*. Processing Pearson Education.

2. Manning, C.D., & Schütze H. (2019). *Foundations of Statistical*. Natural Language Processing MIT.

3. Kim, C., Lee, T., & Kim. H. (2016). Comparison of traditional and deep learning acoustic models for speech recognition in noisy environments. *Electronics Letters*, 52, 10, 822-824.

4. Lee, K., Lee, A., Lee, H., & Kim, S. (2015). Speaker recognition in noisy environments using deep neural networks. *IEEE Signal Processing Letters*, 22, 12, 2339-2343.

5. Gaida, C., & Patrick Lange (2014) Comparing Open-Source Speech Recognition Toolkits. Open Source for Information Systems (OASIS), http://antikenschlacht.de/su/pdf/oasis2014.pdf