**УДК 658.512**

# ДОСЛІДЖЕННЯ МЕТОДІВ РОЗПІЗНАВАННЯ ГОЛОСУ

**Пилипенко В. М.**
Харківський національний університет радіоелектроніки
Україна, 61166, Харків, пр. Науки 14
E-mail: vladyslava.pylypenko@nure.ua

Дослідження проводиться в області розпізнавання мовлення, а саме з використанням нейронних мереж. Метою досліджень у цій області є визначення переваги вибору нейромережевих рішень та безпосередньо рекурентної моделі мережі.

**Ключові слова:** розпізнавання мовлення, нейронна мережа, рекурентна модель.

# MULTIPURPOSE MOBILE ROBOTIC PLATFORM MODELING

**V. Pylypenko**
Kharkiv National University of Radioelectronics
Ukraine, 61166, Kharkiv, Nauky av.,14
E-mail: vladyslava.pylypenko@nure.ua

The research is carried out in the field of speech recognition, namely using neural networks. The aim of research in this area is to determine the advantages of choosing neural network solutions and a recurrent network model.

**Key words:** speech recognition, neural network, recurrent model.

АКТУАЛЬНІСТЬ РОБОТИ. Завдання розпізнавання голосу – одне з найактуальніших завдань сучасності. Незважаючи на те, що на даний момент існує безліч аналогічних готових систем, заснованих на різних технологіях, завдання розпізнавання голосу не повністю вирішене, тому що існуючі системи мають певні недоліки. Зокрема, залежність роботи системи від доступу до засобів передачі даних та недостатня точність розпізнавання.

Одним із перспективних напрямків у вирішенні завдань розпізнавання мовлення є застосування нейронних мереж. Нейронні мережі широко застосовні у вирішенні різних класів завдань розпізнавання через здатність до узагальнення.

АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ РОЗПІЗНАВАННЯ МОВЛЕННЯ. Розпізнавання мовлення – це можливість машини або програми ідентифікувати слова та фрази усною мовою та перетворити їх на машиночитаний формат. Мова являє собою послідовність звуків. Звук, у свою чергу, є суперпозицією звукових хвиль різних частот. Хвиля характеризується двома атрибутами – амплітудою та швидкістю. Щоб зберегти аудіосигнал на цифровому носії, його необхідно поділити на кілька проміжків та прийняти певне «усереднене» значення для кожного з них. Таким чином, механічні коливання перетворюються на набір чисел, які підходять для обробки на сучасних комп'ютерах. Рудиментарне програмне забезпечення для розпізнавання мови має обмежений словниковий запас слів і фраз, і тому воно може ідентифікувати слова, тільки якщо вимова дуже чітка. Більш складне програмне забезпечення може приймати природне мовлення.

Розпізнавання мовлення працює на основі двох алгоритмів: акустичного та мовного моделювання. Акустичне моделювання являє собою взаємозв'язок між лінгвістичними одиницями мови та аудіосигналів; мовне моделювання відповідає звукам із послідовностями слів, щоб допомогти розрізняти слова, які звучать однаково.

Продуктивність систем розпізнавання мови зазвичай оцінюється з точки зору точності та швидкості. Точність оцінюється як кількість помилок у слові, тоді як швидкість вимірюється з коефіцієнтом реального часу. Інші міри точності включають поодиноку помилку та коефіцієнт успіху команди. У процесі розвитку системи розпізнавання мови поступово

з'являлися нові алгоритми роботи, такі як тимчасове динамічне деформування, приховані марківські моделі, нейронні мережі та розпізнавання мови end-to-end.

Одним з ранніх алгоритмів є алгоритм розпізнавання мови на основі динамічного тимчасового деформування (DTW – Dynamic Time Warping). У аналізі часових рядів динамічне тимчасове деформування один із алгоритмів виміру подібності між двома тимчасовими послідовностями. DTW застосовується до тимчасових послідовностей відео-, аудіо- та графічних даних. Дійсно, будь-які дані, які можуть бути перетворені на лінійну послідовність, можуть бути проаналізовані за допомогою DTW, який полягає у вимірі подібності між двома послідовностями, які можуть змінюватися в часі або швидкості.

На зміну алгоритму DTW прийшов більше досконалий підхід – приховані марківські моделі (HMM – Hidden Markov Model). HMM є статистичними моделями, які виводять послідовність символів або величин і використовуються для розпізнавання мовлення, оскільки мовний сигнал можна розглядати як шматково-стаціонарний сигнал або короткочасний стаціонарний сигнал. Кожне слово або фонема має різний розподіл вихідних даних. Фонеми моделюються з використанням трьох різних станів – початкового, середнього та кінцевого. Існує два типи фонем: монофони та трифони. У монофонів накладання артикуляції ігнорується, збираються моделі фонем, що стоять окремо. У трифонів накладання артикуляції враховується, у своїй відбувається побудова окремої моделі для фонем, оточених іншими фонемами. Прихована марківська модель для низки слів або фонем створюється шляхом об'єднання окремих прихованих марківських моделей для кожного слова або фонеми.

Для оптимізації алгоритму HMM часто використовують нейронні мережі, які попередньо обробляють мовний сигнал, наприклад перетворення об'єктів або зменшення розмірності. Штучні нейронні мережі (ANN – Artificial Neural Networks) – це обчислювальні системи, засновані на біологічних нейронних мережах, що становлять мозок тварин. Такі системи вивчають завдання, розглядаючи приклади, як правило, без спеціального програмування. Нейронні мережі є пристроями для зіставлення зразків з архітектурою обробки, заснованої на нейронній структурі людського мозку. Вони складаються з простих взаємозалежних блоків обробки (нейронів). Кожна сполука (синапс) між нейронами може передавати сигнал від одного до іншого. Прийомний (постсинаптичний) нейрон може обробляти сигнал, потім підключати до нього нейрони.

У стандартних реалізаціях ANN синапсовий сигнал є реальним числом, а вихід кожного нейрона обчислюється нелінійною функцією суми його входів. Нейрони та синапси зазвичай мають вагу, яка коригується в міру продовження навчання. Вага збільшує або зменшує силу сигналу, який посилає через синапс. Нейрони можуть мати такий поріг, що тільки в тому випадку, якщо сукупний сигнал перетинає це граничне значення, що посилається сигнал. Як правило, нейрони організовані у шари. Різні шари можуть виконувати різні види перетворень на своїх входи. Сигнали переміщуються від першого (вхідного) до останнього (вихідного) шару. При оцінці ймовірності сегмента мовлення нейронні мережі дозволяють проводити тестування природним та ефективним чином. Недоліком нейронних мереж є нездатність моделювати часові залежності.

Різновидом нейронних мереж є глибокі нейронні мережі (DNN – Deep Neural Network). Даний алгоритм є штучною нейронною мережею з декількома прихованими шарами одиниць між вхідним і вихідним рівнями. Подібно до дрібних нейронних мереж, DNN можуть моделювати складні нелінійні відносини. Архітектури DNN створюють композиційні моделі, в яких додаткові шари дозволяють складати елементи з нижніх шарів, забезпечуючи величезну навчальну здатність і, отже, потенціал моделювання складних моделей мовних даних. DNN мережа має вхідний шар x, прихований шар s та вихідний шар y. Вхідний шар складається з вектора x(t), який є об'єднанням вектора w(t), що є поточним словом, і вектора s(t – 1), який являє собою вихідні значення прихованого шару, отримані на попередньому кроці. Розмір вектора w(t) дорівнює розміру словника. Вихідний шар y(t) має

той же розмір, що і w(t), і після вивчення нейронної мережі являє собою імовірнісне розподіл наступного слова при даному попередньому слові та стан прихованого шару в попередній тимчасовий крок. Розмір прихованого шару зазвичай вибирається емпірично.

Сьогодні найсучаснішим алгоритмом є алгоритм End-to-End пошуку ймовірності зростання, званий LAS (Likelihood Ascent Search). LAS – це модель розпізнавання мови від кінця до кінця. LAS вчиться транскрибувати аудіопослідовність сигналу до послідовності слів, по одному символу за раз, без використання явних мовних моделей, таких як HMM. Він складається з енкодера, який називається listener, і декодера, який називається speller. LAS моделює кожен вихід символу як умовне розподілення порівняно з попереднім символом. Дана модель є дискримінуючою та наскрізною, оскільки вона безпосередньо передбачає умовну ймовірність послідовності символів, враховуючи акустичний сигнал

Головною перевагою систем розпізнавання промови стала дружність до користувача. Вони дозволяють вводити дані або команди через мовлення без використання сенсорних або інших методів. Нестача ж полягає у нездатності розпізнавати деякі варіації вимови, а також відсутність підтримки більшості мов за межами англійської мови та неможливості сортувати фоновий шум. Такі фактори можуть призвести до неточностей [1].

Підсумовуючи, слід зазначити, що, хоча система розпізнавання мови вже розвивається давно, її не можна назвати досконалою, оскільки вона має обмежений потенціал через свою тривіальність. Хоча автоматичні системи розпізнавання мови далеко не ідеальні з точки зору точності слова або завдання, належним чином розроблені програми все ще можуть ефективно використати існуючу технологію для надання реальної цінності клієнту, про що свідчить кількість таких систем, які щодня використовуються мільйонами користувачів. Розвиток систем розпізнавання мови можна пов'язати з удосконаленням структури нейронних мереж, тому було обрано саме цей метод розпізнавання мовлення.

НЕЙРОМЕРЕЖЕВИЙ ПІДХІД. Нейронна мережа з кількістю прихованих шарів є універсальним апроксиматором, тобто навіть мережі з одним прихованим шаром, що використовувалися до цього етапу, можуть апроксимувати будь-яку поверхню у просторі ознак. Проте успіх у розпізнаванні мови прийшов лише з використанням багатошарових мереж. Це пояснюється неможливістю або крайньою складністю створення розумної методики ініціалізації ваг для мереж з одним прихованим шаром, що призводить до далекого від оптимуму набору ваг під час навчання [2].

Одним із методів є ініціалізація за допомогою пошарового навчання, починаючи з нижніх шарів. Як цільову функцію для першого прихованого шару розглядається вхідний вектор ознак. Щоб уникнути тотожного перетворення, вхідний вектор зашумлюють. Наступний шар нейронної мережі таким чином навчають відтворювати вихідні сигнали попереднього шару.

Усього в такий спосіб навчають до 5–7 шарів. Після того, як ініціалізація перших шарів проведена, включають стандартний алгоритм зворотного розповсюдження помилки для всієї мережі з цільовою функцією, що відображає належність вхідного сигналу до трифону. Даний підхід показав явну перевагу в порівнянні з класичним підходом з гаусовими сумішами: результати розпізнавання завжди виявлялися кращими, причому багатошарова мережа, навчена на мовному матеріалі в 309 годин мови, показала кращі результати, ніж метод з гаусовими сумішами, навчений на 2000 годинах мови.

Запропонований алгоритм навчання створює систему, що нагадує функціонування слухову. У слуховій системі виявлено нейрони, які реагують на певні події в акустичному сигналі. У міру «поглиблення» сигналу в центральні відділи слухової системи характер ознак, що виділяються спеціалізованими нейронами, набуває все більш складного та вибіркового характеру. Попереднє навчання окремих шарів нейронної мережі, виконує те саме завдання – окремі шари навчаються знаходити ознаки сигналу дедалі вищого рівня.

Якщо внутрішні шари нейронних мереж виділяють ознаки мовного сигналу, притаманні мовленню взагалі, їх можна уніфікувати для всім мов, навчаючи для кожній новій мові лише вихідний шар нейронної мережі. Це було б надзвичайно важливо, оскільки для навчання

лише одного шару нейронної мережі була б потрібна набагато менша мовна база даних, ніж для навчання всіх 5–7 шарів. Експерименти повністю підтвердили таку можливість.

Оскільки нейронні мережі не можуть ідентифікувати динамічні об'єкти, для порівняння моделей із сигналом, як і раніше, використовується формалізм марківських моделей, проте тепер як вектор ознак використовується набір апостеріорних ймовірностей трифонів, отриманий на виході нейронної мережі. Більш істотним недоліком, властивим даному методу, є те, що глибокі нейронні мережі не можуть розпізнавати динамічні об'єкти, через що й доводиться використовувати алгоритми марківської моделі. Недоліки марковської моделі досить очевидні: дискретність, чи незалежність послідовних станів друг від друга; відсутність глибоких тимчасових зв'язків, тобто нездатність розпізнавати траєкторії у просторі ознак як інформативні об'єкти.

Можна припустити, що обидва зазначені недоліки можна подолати, використовуючи рекурентних нейронні мережі (РНМ). РНМ містять нейрони, які об'єднані в спрямований круговий процес. Це наділяє нейронну мережу пам'яттю і, отже, здатністю розпізнавати процеси, а чи не лише статичні об'єкти, як розглянуті вище глибокі нейронні мережі. РНМ відрізняються від розглянутих раніше багатошарових тим, що з обробці чергового вектора ознак система враховує також внутрішні стани нейронів, які, своєю чергою, формуються попередніми векторами ознак і станами попередні моменти часу. У цьому сенсі одинична рекурентна нейронна мережа є потужнішим освітою, ніж глибока нейронна мережу. Проте, розглядаються ієрархічні комбінації РНМ та комбінації рекурентних та багатошарових мереж. Це можна пояснити, як і для багатошарових мереж, бажанням структурувати систему, наблизити її до принципів функціонування нервової системи, спростити процедуру ініціалізації та навчання.

ВИКОРИСТАННЯ РЕКУРЕНТНИХ НЕЙРОННИХ МЕРЕЖ. Процесом розпізнавання мовлення складається з мікрофона, за допомогою якого люди можуть говорити, програмного забезпечення для розпізнавання мовлення та комп’ютера для виконання завдання. Основне розпізнавання мовної системи показано на рис. 1.

Для перетворення звукових хвиль в текст його необхідно подати в комп'ютер. Оскільки звукові хвилі є безперервним (аналоговим) сигналом, перше, що потрібно зробити – провести дискретизацію сигналу за допомогою теореми Найквіста. Цей дискретизований сигнал надходить безпосередньо в нашу нейронну мережу, але необхідно виконати попередню обробку сигналу, щоб отримати кращий результат і точні прогнози слів, що розпізнаються. Попередня обробка – це угруповання великого дискретизованого сигналу на невеликі фрагменти по 20 мс, як приклад.

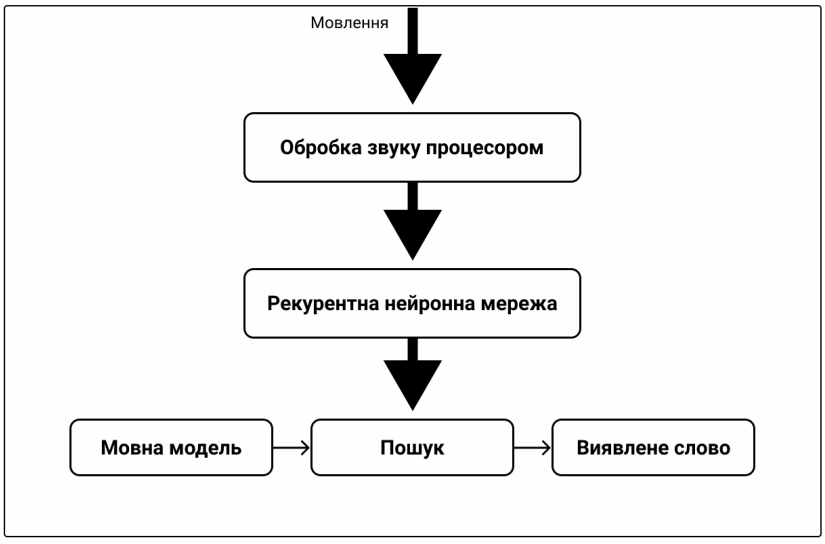


Рисунок 1 – Основне розпізнавання мовної системи

РНМ є основною моделлю розпізнавання мови, що використовується для прогнозування. На рис. 2 показано розпізнавання мовлення з використанням РНМ [3]. Тепер звук, який подається на вхід, легко обробляти, він подаватиметься в глибоку нейронну мережу. Після подачі в мережу невеликих звукових фрагментів тривалістю близько 20 мс, він визначить букву, яка відповідає звуку, що вимовляється.

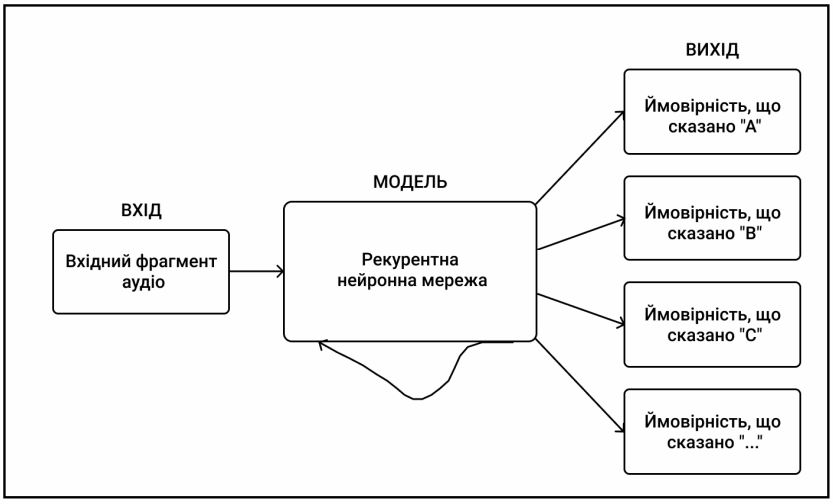


Рисунок 2 – Розпізнавання мовлення з використанням РНМ

RNN – це мережа з пам'яттю, яка визначає майбутні прогнози. Це пов'язано з тим, що, оскільки він передбачає одну літеру, це впливає на ймовірність наступної літеру, яку він також передбачає. Отже, наявність пам'яті про попередні прогнози допомагає мережі робити точніші прогнози в майбутньому.

RNN використовує ідею послідовної інформації, так як ця нейронна мережа має пам'ять, що впливає на майбутні прогнози. Для прогнозів використовується послідовна інформація, що зберігається у пам'яті RNN. Ідея використовувати RNN замість традиційної нейронної мережі полягає в тому, що в традиційній нейронній мережі передбачається, що кожен вхід і кожен результат залежать друг від друга. Отже, використання традиційної нейронної мережі – погана ідея при обробці мови. Передбачення будь-яких слів у реченні вимагає інформації про слово, що використовується до того, як минуле слово обробляється. Наявність пам'яті – одна з особливостей RNN, що робить її унікальною проти іншими мережами. Отже, RNN є найефективнішою для розпізнавання мови.

ЛІТЕРАТУРА

1. Белов Ю. С., Либеров Р. В. Подходы и проблемы распознавания личности по голосу // Электронный журнал: наука, техника и образование. 2015. № 3 (3). С. 68–77.
2. Применение искусственных нейронных сетей для распознавания речевых команд / Г. К. Бердибаева, О. Н. Бодин, Н. В. Громков, В. В. Козлов, К. А. Ожикенов, Я. А. Пижонков // Измерение. Мониторинг. Управление. Контроль. 2017. № 2 (20). С. 77–84.systems. Innovative Technologies and Scientific Solutions for Industries, (4 (14)), 155–168.
3. Aditya Amberkar, Parikshit Awasarmol, Gaurav Deshmukh, Piyush Dave. (2018). Speech Recognition using Recurrent Neural Networks. Conference: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). DOI:10.1109/ICCTCT.2018.8551185

***Науковий керівник:*** *Євсєєв Владислав В'ячеславович, д.т.н., професор кафедри КІТАМ Харківського національного університету радіоелектроніки*