УДК 004.522

А.Ю. Романенко, В.В. Олійник

УЗАГАЛЬНЕНА МОДЕЛЬ РОЗПІЗНАВАННЯ ГОЛОСОВИХ КОМАНД

Abstract: the principles of control systems' optimal design are given while it is running two main modes: stationary and automatic. The search methods of decisions in accordance with these modes are offered.

Анотація: Представлена модель системи розпізнавання голосових команд спрощує розробку нових практичних реалізацій систем розпізнавання команд, полегшує модифікацію існуючих, спрощує порівняння ефективності різних алгоритмів розпізнавання мови за рахунок модульної будови, де кожен блок виконує важливу функцію з точки зору вищеназваної задачі розпізнавання.

Ключові слова: розпізнавання мови, голосове управління, модель розпізнавання, виділення ознак, розпізнавання фонем, розпізнавання слова.

Вступ

Системи розпізнавання усної мови мають дуже високий потенціал для практичного застосування і тому стрімко розвиваються в сьогоденні. Згідно із прогнозами агентства Tractica, очікується стрімкий ріст ринку технологій машинного розпізнавання мови від 70млн. доларів у 2016р до понад 500млн. доларів у 2024 році[1]. Переваги цих систем полягають у швидкості та зручності вводу інформації в інформаційну систему з її одночасною інтерпретацією у внутрішнє подання бази знань чи збереженням у більш загальній текстовій формі з можливістю подальшої обробки за допомогою існуючих засобів.

Розпізнавання мови – напрям, який поєднує багато складових, основні з яких: розпізнавання спонтанної мови, неперервної мови, поєднаних слів, ізольованих слів, тощо. А розпізнавання голосових команд – є однією з найважливіших з точки зори практичного застосування підзадачею розпізнавання мови, оскільки є фундаментом для загальної задачі розпізнавання мови та знаходить свій подальший розвиток у задачі розпізнавання неперервного мовлення.

Незважаючи на дуже велику кількість існуючих практичних та теоретичних рішень цієї задачі, досі активно продовжуються пошуки підходів, які б збільшили точність та швидкість розв'язання; зменшили кількість та строгість обмежень, що накладаються на умови задачі [2].

Крім того, існуючі прикладні системи розпізнавання команд мають низку обмежень для застосування (кількість запитів на день, час відповіді, залежність від наявності інтернету тощо), вирішення яких призводить до ще більшого зростання числа подібних підходів, моделей та реалізованих систем. Аналіз та порівняння цих моделей і систем з метою практичного використання та синтезу нових систем вимагають глибоких знань предметної області та суттєво ускладняються браком інформації про подробиці реалізації прикладних систем, що створює високий поріг входження для інженерів, які займаються проблемами розпізнавання голосових команд.

Метою даної статті є створення узагальненої моделі системи розпізнавання голосових команд, яка мала б модульну структуру та слугувала б перехідною ланкою від абстрактних до спеціалізованих моделей. Таким чином, елементи існуючих прикладних рішень та теоретичних моделей можна буде співвідносити з елементами запропонованої у статті моделі. У результаті такої уніфікації синтез нових та порівняння існуючих практичних рішень і їх елементів значно спроститься, що зменшить поріг входження та прискорить розробку нових систем розпізнавання голосових команд (аналогічно підходу до створення семантично-узгодженого середовища прийняття рішень [6]).

Задача розпізнавання голосових команд

Під задачею розпізнавання голосових команд мається на увазі перетворення звукової хвилі в команду, де командою є одне слово, рідше словосполучення. Звукова хвиля спочатку та вкінці повинна бути обмежена "тишею” - проміжками, коли немає мовлення.

Таким чином, задача перетворення усних команд в текст є задачею розпізнавання образів. При цьому сигнал, що розпізнається, у випадку мови, є розподіленим у часі. Це означає, що у процесі розпізнавання весь час надходить нова інформація, яка може впливати на попередні результати роботи алгоритму.

Класично задача розпізнавання образів розв’язується у 2 етапи:

1) Виділити із вхідного сигналу множину репрезентативних ознак.

2) Використовуючи виділені ознаки, віднести вхідний сигнал до одного із класів за певним алгоритмом.

Проте у випадку мови було показано, що такий прямий підхід не є ефективним. Це відбувається тому, що слово є комбінацією простіших

одиниць – фонем. Кожна фонема має свій набір характеристик, які можуть змінюватись. Тому важко чи навіть практично неможливо розробити такий набір ознак, які б характеризували слово з урахуванням характеристик всіх його фонем та, відповідно, комбінацій характеристик фонем. Внаслідок цього практично всі успішні системи є ієрархічними: вони розв'язують дві (або більше) задачі розпізнавання.

Передумови створення моделі системи розпізнавання голосових команд

На сьогоднішній день розпізнавання мови - це напрям, який продовжує активно розвиватись. Відповідно, створюються промислові системи на зразок Google Cloud Speech API, які зберігають дані про будову системи розпізнавання мови як промислову таємницю. Є велика кількість наукових публікацій щодо покращення якості (за критеріями точності та швидкості) розв'язку різних підзадач із задачі розпізнавання голосових команд. Також в публікаціях задача розпізнавання мови розглядається в цілому. Існують системи та проекти з відкритим кодом, у яких розв'язок задачі описаний із найменшими подробицями у коді.

Одним із недоліків існуючих моделей систем розпізнавання голосових команд є занадто велика спеціалізація, коли є модель конкретного прикладного оптимізованого рішення з інфраструктурою. У цьому випадку модель системи занадто обтяжена деталями реалізації, що приховують за собою основну модель розпізнавання мови; також при оптимізації кілька компонентів можуть бути злиті в один, що ще більше ускладнює розуміння моделі. Моделі у наукових статтях зазвичай концентруються на одній спеціалізованій задачі, при цьому опускається або приділяється мало уваги тому, як їхнє рішення впливає та може вписатись у вже існуючі системи.

Іншим недоліком моделей є занадто велика абстрактність. Зазвичай такі моделі оперують високорівневими елементами і припускають, що ці елементи вже розв'язують задачі на зразок розпізнавання фонем та розпізнавання слів. Часто вони зустрічаються у підручниках чи загальних статтях про задачу розпізнавання мови. У цьому разі необхідні глибокі знання предметної області, щоб замінити абстрактні елементи системи на конкретні для розв'язку поставленої задачі.

За такого стану речей інженери, які вирішують побудувати власну систему розпізнавання голосових команд стикаються із високим поро-

гом входження. Для початку їм необхідно розглянути велику кількість публікацій, розібратися в математичних моделях, переглянути реалізації існуючих систем, визначити на діаграмах класів ядро системи розпізнавання команд. Також вони мають визначити області застосування та співставити межі задач, які вирішують елементи існуючих систем та моделі і алгоритми описані в публікаціях. На основі проведеного аналізу визначити можливі комбінації цих рішень та відповідно структуру системи, яка вийде у результаті. Для людини без спеціальних знань цей процес є дуже складним, оскільки вимагає прикладення великої кількості зусиль та займає багато часу.

Задачею даної статті є впорядкування існуючих методів та підходів у межах задачі розпізнавання голосових команд у вигляді узагальненої моделі. Створення системи розпізнавання голосових команд на базі цієї моделі проходить швидше, бо не потрібно розбиратися в деталях та особливостях роботи модулів, можна використовувати їх як “чорні ящики”. З іншого боку, полегшується модифікація існуючих систем, створених на базі цієї моделі: кожен елемент має своє місце, відповідно, можна замінити частину системи та дослідити як це впливає на ефективність її роботи.

Модель системи розпізнавання голосових команд

Представлена модель вирішує задачу розпізнавання голосових команд у 2 етапи, кожен з яких є задачею розпізнавання образів: 1 етап - співставлення звукового фрагмента із кількома класами фонем; 2 - етап співставлення класів фонем із словами з словника для визначення слова. Вона має модульну структуру, де кожен модуль вирішує невелику, але вагому підзадачу для задачі розпізнавання голосових команд.

Таке рішення робить модель більш прозорою: вона не обтяжена деталями інфраструктури, а показує лише важливі для розпізнавання голосових команд кроки. Іншою перевагою є достатня деталізація в сенсі того, що кожному модулю системи відповідає певний алгоритм обробки звукової інформації. Це означає, що інженери можуть обрати інтерфейси для передачі даних між блоками, застосувати уже готові алгоритми та отримати робочий прототип системи.

Розроблена модель системи розпізнавання голосових команд зображена на рис. 1. Модель системи складається з блоків:

Зчитування звукової хвилі є першим етапом у роботі системи - це задача тривіальна і вирішена ще у минулому столітті. Сучасні пристрої

для звукозапису дозволяють отримувати цифровий сигнал дуже високої якості, з частотою дискретизації більше 44кГц. Коли для повноцінної передачі мови вистачає смуги частот у 4кГц.

Реалізація: Популярні мови програмування мають бібліотеки, які дозволяють зчитувати звуковий сигнал із файлів та пристроїв звукозапису.

Фільтрація. Сегментація - цей блок призначений для виділення корисного сигналу та подавлення шумів, а також виділенням мови від тиші (сегментації). У випадку розпізнавання неперервної мови цей блок не займається виділенням окремих слів, бо під час мовлення слова можуть бути відсутні паузи між словами.

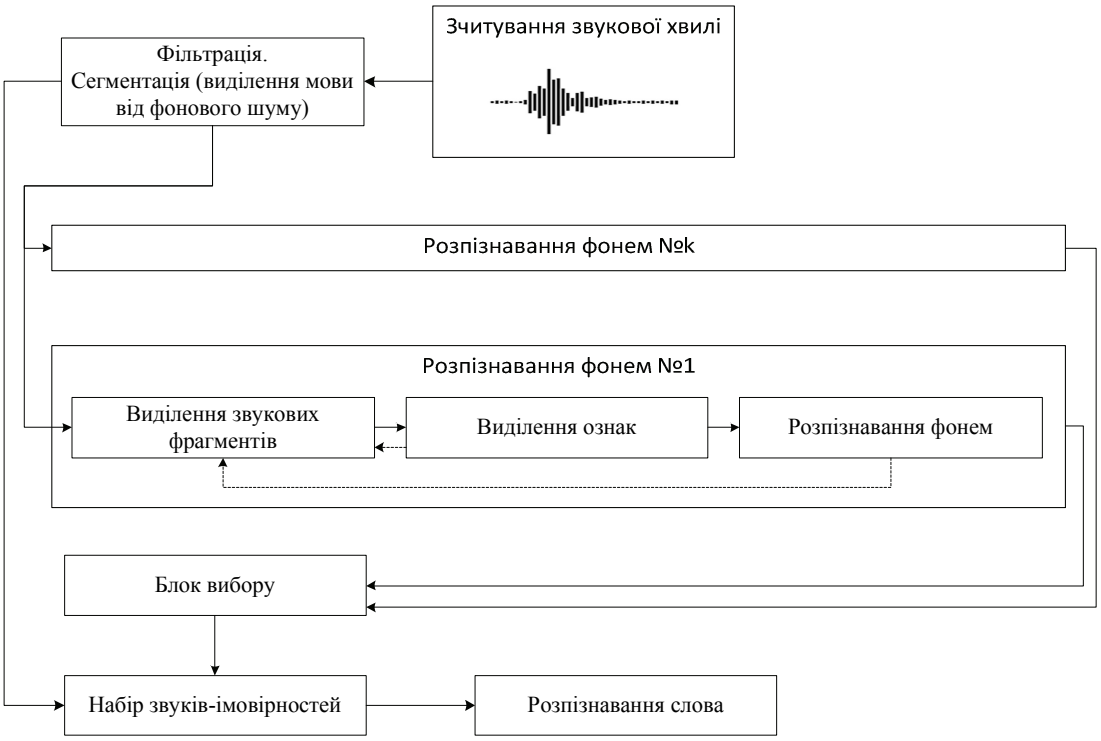

Рис. 1. Модель системи розпізнавання голосових команд

Складність у даному випадку полягає у тому, що важко розробити фільтри та алгоритми сегментації, які б працювали однаково якісно завжди незалежно від умов. Але знаючи природу шумів можна розробити ефективні алгоритми фільтрації та сегментації.

Реалізація: У найпростішому випадку можна використати цифрові фільтри, а для сегментації виділення мови за показником енергії та кількості перетинів нуля (Zero Crossing Rate).

Розпізнавання фонем - на цю підсистему покладено найбільш відповідальну роботу. Вона перетворює звуковий сигнал у послідовність

фонем. Від точності роботи цієї системи залежить об'єм словника та точність розпізнавання в цілому.

Суть задачі полягає в тому, що потрібно перевести сигнал, розтягнутий у часі, у набір статичних ознак, які не залежать від часу (в ідеальному випадку). Одна і та ж фонема може давати різний часовий ряд, залежно від мовця та фонем, що стоять поруч. А значення набору ознак, які представляють ці ряди, повинні бути дуже схожими, у ідеальному випадку - майже однаковими.

На практиці досягнути цього неможливо, бо поряд з вище описаними складнощами, на сигнал кожного разу впливають випадкові шуми з невідомими статистичними характеристиками (зокрема сам голосовий тракт людини може створювати ці шуми).

У найпростішому варіанті результатом роботи цієї підсистеми є пара значень:

$$R = \{z; p\}, \tag{1}$$

де z – фонема із множини допустимих фонем , яка відповідає звуку на вході. Найчастіше множина допустимих фонем складається зі звуків мови, для якої будується система (наприклад, для української мови $Z = \{a, \varepsilon, i, u, ɔ, ɪ, m, n, dz, dʒ, \dots\}$); p – імовірність того, що вхідний сигнал є фонемою z.

У загальному варіанті результатом роботи підсистеми є набір кортежів:

$$R = (\{z_1; p_1\}, \dots \{z_i; p_i\} \dots, \{z_k; p_k\}), \tag{2}$$

де k – кількість пар, які співставляють кожному звуковому фрагменту; $i = 1..k$ – номер пари; z_i - фонема із множини допустимих фонем , яка відповідає звуку на вході; p_i - імовірність того, що вхідний сигнал є фонемою z_i.

Для різних звуків окремі алгоритми виділення ознак та розпізнавання мають більшу ефективність, ніж інші. Тому у системі може існувати більше однієї підсистеми розпізнавання фонем.

Реалізація: Підсистема розпізнавання фонем складається із кількох простіших підсистем, які взаємодіють між собою: виділення звукових фрагментів, виділення ознак звукового фрагмента, розпізнавання фонем.

Блок виділення звукових фрагментів відповідає за те, щоб у неперервному потоці виділити окремі звукові фрагменти. Вона повідомляє

про початок нового звукового фрагмента та його кінець підсистемі виділення ознак звуку. Виділити одну морфему від іншої - нетривіальна та складна задача, тому цей блок може додатково використовувати результати роботи блоків виділення ознак та розпізнавання звуків.

Найпростіший та найбільш вживаний варіант рішення цієї задачі - перемножувати вхідний сигнал із віконною функцією фіксованої довжини. При цьому вважається, що у вікно потрапляє не більше, ніж одна фонема. Відповідно до цього обирається довжина вікна (зазвичай 10-20мс вхідного сигналу). Віконна функція при цьому рухається із перекриттям, на кожному кроці відбувається зсув на 10-50% ширини вікна:

$$x(n) = y(n)w(n - t) = y(n)w(n - i \cdot k), \qquad (3)$$

де $x(n)$ – звуковий фрагмент, виділений за допомогою віконної функції; $y(n)$ – вхідний звуковий сигнал; $w(n)$ – віконна функція (прямокутне вікно, вікно Хеммінга, тощо); t – поточний зсув вікна; k – величина зсуву; i – номер поточного фрагмента.

Такий підхід має в собі велику надмірність, але він простий у реалізації і дозволяє виділити частину сигналу, що відповідає фонемі із високою точністю.

Блок виділення ознак звукового фрагмента перетворює звуковий фрагмент у набір ознак, які подаються у блок розпізнавання фонем. Основна задача у цьому випадку перетворити виділений звук на вході у вектор чи число, яке описуватиме його як фонему. Виділення репрезентативних ознак є нетривіальною задачею, як наслідок існує багато підходів до її вирішення. Кожен має свої переваги, недоліки та область застосування. Багато методів базується на інтуїтивному припущенні, що ці ознаки повинні враховувати частотну інформацію, яка міститься в сигналі, за аналогією до того, як людське вухо обробляє звук.

Найбільш поширеними методами є виділення ознак за допомогою *віконного перетворення Фур'є, мел-кепстральних частотних коефіцієнтів, банку фільтрів, вейвлет перетворення, коефіцієнтів лінійного передбачення та використання чистого сигналу.*

Блок розпізнавання фонем приймає на вхід ознаки поточного фрагмента звуку та повертає імовірність того, що даний фрагмент є певною фонемою $R = \{z; p\}$ або набір імовірностей фонем $R = (\{z_1; p_1\}, \ldots, \{z_k; p_k\})$. Цей блок разом із попереднім є ядром системи розпізнавання команд. Від точності та якості їх роботи залежить точність і якість роботи системи

 ISSN 1560-8956

вцілому. Для розпізнавання частіше за все використовують некласичні алгоритми розпізнавання образів - алгоритми на основі машинного навчання. Причиною цьому є те, що форма функцій, що відділяють класи є надзвичайно складними. Крім того часто неможливо чітко провести розділяючу поверхню і доводиться враховувати, що один і той самий фрагмент може бути однією або іншою фонемою. Вирішити яка саме це була фонема може блок розпізнавання слова.

Щоб передавати більш точну та повну інформацію до блоку розпізнавання слова, даний блок може повертати не фонеми, а алофони – варіант реалізації фонеми, зумовлений її оточенням.

Для розпізнавання фонем використовуються *приховані моделі Маркова, нейромережі (згорткові та рекурсивні)* рідше алгоритми *динамічного програмування на зразок DTW*.

Блок вибору розглядає результати роботи підсистем або підсистеми розпізнавання фонем та обирає найбільш імовірні фонеми, які відповідають звуковому фрагменту на вході. Для кожного фрагмента це може бути набір із кількох фонем та імовірностей $R = (\{z_1 : p_1\} \dots, \{z_k : p_k\})$, щоб дати блокові розпізнавання слів більше інформації, підвищивши точність розпізнавання.

Реалізація: найбільш простим, але не менш дієвим критерієм для вибору наборів є імовірність. Найпростіша реалізація блоку вибору обирає задану кількість фонем, які мають найбільшу імовірність. Більш досконалим є варіант, коли імовірність ще множиться на коефіцієнт, що залежить від попередніх результатів розпізнавання. Наприклад, якщо попередній звук був приголосним, то імовірності всіх приголосних у цьому фрагменті множаться на 0.9, якщо 2 попередні приголосні – 0.6 і т. д.

Набір звуків-імовірностей даний блок зберігає послідовність розпізнаних фонем $W = (R_1 \dots, R_j)$. Важливо, щоб порядок наборів R_j зберігались у тій же послідовності, в якій вони проходять обробку. У подальшому вектор W служить набором ознак для блоку розпізнавання слова. Блок сегментації надає інформацію цьому блоку про те, коли почалось і закінчилось слово. Таким чином, після закінчення слова вектор W подається далі на розпізнавання.

Блок розпізнавання слова виконує останню частину роботи - перетворення масивів фонем та імовірностей у слово. У зв'язку з тим, що підсистема розпізнавання фонем доволі неточна, цей блок повинен ком-

пенсувати похибки її роботи. У загальному випадку залежно від кількості пар фонем-імовірностей, що співставляється з одним фрагментом, може бути різна кількість інформації, яку треба обробити на цьому етапі. Найбільш дієвою допомогою у цьому випадку є використання словника, що дозволяє суттєво зменшити кількість варіантів для перебору. Чим менший розмір словника, тим легше скорегувати похибки роботи попередніх блоків і тим простішими можуть бути алгоритми розпізнавання. З ростом розмірів словника відповідно підвищуються вимоги до точності роботи цього блоку і, як результат, складніші алгоритми повинні бути використані.

У даному блоці застосовуються методи такі *як мінімум відстані, динамічне програмування (найдовша спільна послідовність), приховані моделі Маркова* та алгоритми *на основі скінченних автоматів*.

Висновки

У даній статті описано недоліки існуючих моделей систем розпізнавання мови та голосових команд і запропоновано нову модель, яка позбавлена цих недоліків. Будова запропонованої моделі детально описана, що дає можливості для її застосування при розробці нових систем розпізнавання голосових команд. Завдяки тому, що модель має модульну структуру, зменшується об'єм інженерних робіт, необхідних для перевірки ефективності роботи нових алгоритмів для конкретних блоків. Вказані алгоритми, що можуть використовуватись у блоках в межах запропонованої моделі, що зменшує поріг входження та коло пошуку для інженерів, що починають розробку власної системи. Ефективність та доцільність використання окремих алгоритмів-складових моделі буде описана в наступних роботах.

Список використаних джерел

1. Voice and Speech Recognition Software Licenses to Surpass 550 Million Worldwide by 2024 | Tractica [Електронний ресурс] – Електронні дані. – Режим доступу: https://www.tractica.com/newsroom/press-releases/voice-and-speech-recognition-software-licenses-to-surpass-550-million-worldwide-by-2024/ – Дата доступу: 12.04.17 – Назва з екрана.

2. Lawrence R. B. Fundamentals of speech recognition / Lawrence Rabiner Biing, Hwang Juang – Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 1993 – 507 с.

3. Распознавание речи от Яндекса. Под капотом у Yandex.SpeechKit [Електронний ресурс] – Електронні дані. – Режим доступу: https://habrahabr.ru/company/yandex/blog/198556/ – Дата доступу: 12.04.17 – Назва з екрана.

4. EE E6820: Speech & Audio Processing & Recognition [Електронний ресурс] – Електронні дані. – Режим доступу: https://www.cs.ubc.ca/~murphyk/Software/HMM/E6820-L10-ASR-seq.pdf – Дата доступу: 12.04.17 – Назва з екрана.

5. Learning Acoustic Frame Labeling For Speech Recognition With Recurrent Neural Networks / Hasjm Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Franc oise Beaufays, Johan Schalkwyk // Google – Електронні дані. – Режим доступу: https://static.googleusercontent.com/media/research.google.com/uk//pubs/archive/43908.pdf – Дата доступу: 12.04.17 – Назва з екрана.

6. Олейник В.В. Рациональный выбор формализмов семантически-согласованной среды при моделировании компьютерно-интегрированных производственных систем / В.В. Олейник, Л.С. Ямпольский, О.И. Лисовиченко // Адаптивні системи автоматичного управління: міжвід. наук.- тех. збірник. – Дніпропетровськ: ДНВП Системні технології. – 2006. – №9(29). – С. 93-101.