УДК 004.934 DOI https://doi.org/10.30837/IYF.PDICIMT.2024.187

ОЦІНКА ЯКОСТІ РОЗПІЗНАВАННЯ ГОЛОСОВИХ КОМАНД ЛЮДИНИ

Шишаков Є.В.
Науковий керівник – доц., к.т.н. Омельченко С.В.
Харківський національний університет радіоелектроніки, каф. IMI,
м. Харків, Україна
E-mail: yevhenii.shyshakov@nure.ua.

This work is devoted to assessing the field of artificial neural networks has grown rapidly in recent years. This has been accompanied by an insurgence of work in speech recognition. Most speech recognition research has centered on stochastic models, in particular the use of hidden Markov models (HMMs). Alternate techniques have focused on applying neural networks to classify speech signals. The inspiration for using neural networks as a classifier stems from the fact that neural networks within the human brain are used for speech recognition. This analogy unfortunately falls short of being close to an actual model of the brain, but the modeling mechanism and the training procedures allow the possiblility of using a neural network as a stochastic model that can be discrimitively trained.

Автоматичне розпізнавання мовлення (ASR), яке спрямоване на природну взаємодію між людиною та машиною, було предметом інтенсивних досліджень протягом десятиліть. Багато основних технологій, таких як моделі суміші Гауса (GMM), приховані моделі Маркова (HMM), кепстральні коефіцієнти мел-частоти (MFCC) та їх похідні, моделі мови ngram (LM), дискримінаційне навчання та різні методи адаптації, були розроблені разом До речі, переважно до нового тисячоліття. Ці методи значно просунули сучасний рівень ASR та суміжних галузей. Порівняно з цими попередніми досягненнями, прогрес у дослідженні та застосуванні ASR у десятиліття до 2010 року [1] був відносно повільним і менш захоплюючим, хоча важливі методи, такі як розрізнювальна підготовка послідовності GMM–HMM, у цей період добре працювали в практичних системах.

Однак за останні кілька років ми спостерігаємо новий сплеск інтересу до ASR. На нашу думку, ця зміна була спричинена підвищеними вимогами до ASR у мобільних пристроях і успіхом нових мовних програм у мобільному світі, таких як голосовий пошук (VS), диктування коротких повідомлень (SMD) і віртуальні мовні помічники (наприклад, Siri від Apple, Google Now і Cortana від Microsoft). Не менш важливою є розробка методів глибокого навчання [2][4][5] в системі безперервного розпізнавання мовлення великого словника (LVCSR), що базується на великих даних і значно покращує обчислювальну здатність. Поєднання набору методів глибокого навчання призвело до зниження частоти

помилок більш ніж на 1/3 у порівнянні зі звичайною сучасною структурою GMM–HMM для багатьох реальних завдань LVCSR і допомогло подолати поріг прийняття для багатьох користувачів реального світу. Наприклад, точність слова в англійській мові або точність символів в китайській мові в більшості систем SMD зараз перевищує 90 %, а в деяких системах навіть 95 % [3].

У розпізнаванні мовлення один із найпоширеніших генеративних підходів до навчання базується на прихованих моделях Маркова на основі моделі суміші Гауса, або GMM-HMM [1]. Як обговорювалося раніше, GMM-HMM — це статистична модель, яка описує два залежні випадкові процеси, спостережуваний процес і прихований процес Маркова. Передбачається, що послідовність спостереження генерується кожним прихованим станом відповідно до розподілу суміші Гауса. GMM-HMM параметризується вектором попередніх ймовірностей стану, матрицею ймовірностей переходу стану та набором залежних від стану параметрів у моделях суміші Гаусса. З точки зору моделювання мовлення, стан у GMM-HMM зазвичай асоціюється з підсегментом телефону в мовленні [3].

Одним з важливих нововведень у використанні HMM для розпізнавання мовлення є введення контекстно-залежних станів, мотивоване бажанням зменшити варіабельність вихідних векторів ознак мови, пов’язаних з кожним станом, загальною стратегією для «детального» генеративного моделювання. Наслідком використання залежності від контексту є значне розширення простору станів HMM, яким, на щастя, можна керувати методами регулярізації, такими як зв’язування станів.

Виявляється, така залежність від контексту також відіграє вирішальну роль у нещодавньому прогресі розпізнавання мовлення в області глибокого навчання на основі дискримінації.

Запровадження HMM та відповідних статистичних методів для розпізнавання мовлення в середині 1970-х років можна вважати найбільш значущою зміною парадигми в галузі, як обговорювалося та аналізувалося в. Однією з головних причин такого раннього успіху є високоефективний алгоритм EM. Цей метод максимальної правдоподібності, який часто називають алгоритмом Баума-Велча, був основним способом навчання систем розпізнавання мовлення на основі HMM до 2002 року, і досі є одним із основних кроків (серед багатьох) у навчанні цих систем сьогодні. Цікаво відзначити, що алгоритм Баума-Велча служить одним з головних мотивуючих прикладів для подальшого розвитку більш загального алгоритму EM.

Використання генеративної моделі HMM для представлення (порізно стаціонарного) динамічного шаблону мовлення та використання EM-алгоритму для навчання пов’язаних параметрів HMM є одним з найвидатніших і успішних прикладів генеративного навчання в розпізнаванні мовлення [1]. Цей успіх був міцно закріплений мовленнєвою

спільнотою та широко поширений на машинне навчання та пов’язані спільноти. Насправді HMM став стандартним інструментом не лише для розпізнавання мовлення, але й для машинного навчання, а також у суміжних областях, таких як біоінформатика та обробка природної мови. Для багатьох дослідників машинного навчання та розпізнавання мовлення успіх HMM у розпізнаванні мовлення є дещо дивним через добре відомі недоліки HMM у моделюванні динаміки мовлення.

Список використаних джерел:

1. Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods, by Joseph Keshet, Samy Bengio (January 2009)
2. Speech Recognition Over Digital Channels: Robustness and Standards, by Antonio Peinado and Jose Segura (September 2006)
3. Speech Processing — A Dynamic and Optimization-Oriented Approach, by Li Deng and Doug O’Shaughnessy (June 2003)
4. Digital Speech Processing: Synthesis, and Recognition, Second Edition, by Sadaoki Furui (June 2001)
5. Speech Communications: Human and Machine, Second Edition, by Douglas O’Shaughnessy (June 2000)