УДК 621.391.19

Р.Б. Попович
Національний університет “Львівська політехніка”,
кафедра ЕОМ

СКЛАДОВІ СИСТЕМИ РОЗПІЗНАВАННЯ МОВИ

© Попович Р.Б., 2001

Проаналізовано складові сучасних систем розпізнавання суцільної мови з великим словником та завдання, які виникають при реалізації цих складових.

One has analized components of current large vocabulary continuous speech recognition systems and problems arising in case of these components realization.

Сучасні системи для розпізнавання суцільної мови з великим словником грунтуються на принципах статистичного розпізнавання образів [1-5]. На рис. 1 показана структура типової системи статистичного розпізнавання.

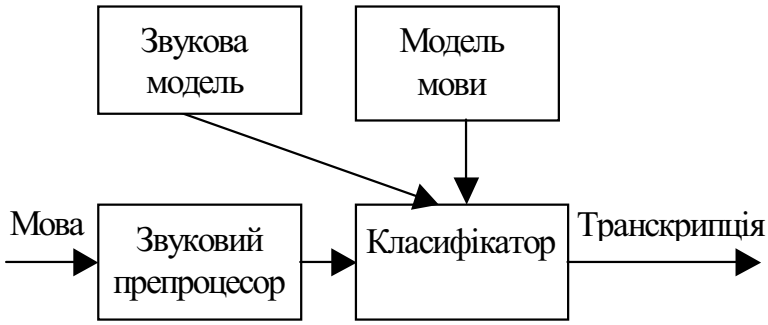

Рис. 1. Структура системи розпізнавання мови

Розглянемо кожну з її чотирьох складових більш детально.

Звуковий препроцесор. Потрібен початковий етап обробки, на якому з мовного сигналу вибирається вся необхідна звукова інформація в компактному вигляді.

Принципове припущення, яке робиться в сучасних розпізнавачах [1, 4] – це те, що мовний сигнал розглядається як стаціонарний (тобто спектральні характеристики відносно постійні) на інтервалі в кілька десятків мілісекунд. Тому основною функцією попередньої обробки є розбиття вхідної мови на інтервали [1,4] і отримання для кожного інтервалу згладженої спектральної оцінки. Зсув між інтервалами зазвичай дорівнює 10 мс. Інтервали, як правило , перекриваються і мають тривалість 25 мс. Переважно для обробки такого типу до кожного інтервалу на початку застосовується функція вікна (наприклад, вікно Хемінга). Часто застосовують високочастотне підсилення, щоби компенсувати послаблення, спричинене розсіюванням від губ.

Щоб отримати спектральні оцінки, використовується швидке перетворення Фур’є.

Фур’є-спектр згладжується додаванням спектральних коефіцієнтів у межах “трикутних” частотних смуг розташованих на нелінійній (подібній до логарифмічної) Mel-шкалі [4]. Для граничної частоти мови, що дорівнює 8 КГц, беруть 24 такі частотні смуги. Mel-шкала введена для наближення частотного розділення людського вуха, яке є лінійним до 1000 Гц та логарифмічним понад 1000 Гц.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Щоби зробити статистику оціненого спектру потужності мови близькою до гауссової, до виходів набору фільтрів застосовують логарифмічний стиск.

До прологарифмованих коефіцієнтів застосовують дискретне косинусне перетворення. Це зосереджує спектральну інформацію в кепстральних коефіцієнтах з малими номерами, а також декорелює їх, дозволяючи при подальшому статистичному моделюванні використовувати діагональні коваріаційні матриці. Перші 12 кепстральні коефіцієнти та логарифм енергії інтервалу сигналу утворюють базовий 13-елементний звуковий вектор.

Є ряд додаткових перетворень, які можна застосувати для отримання остаточного звукового вектора.

Для зменшення мультиплікативного шуму на звукових векторах нормалізуються кепстральні коефіцієнти. Для кожної з дванадцяти компонент обчислюються середні значення за всіма звуковими векторами даного мовного зразка. Ці середні значення віднімаються від відповідних компонент всіх звукових векторів даного мовного зразка.

Компоненти логарифма енергії даного мовного зразка, які менші від максимального значення на 50 дБ, заміняються на значення цього порогу. Потім всі значення логарифма енергії масштабуються так, що максимальне значення дорівнює 1,0.

Припускається, що кожен звуковий вектор не зв’язаний зі своїми сусідами. Це досить грубе припущеня, бо фізичні обмеження голосового тракту людини передбачають плавні переходи між сусідніми спектральними оцінками. Проте додавання різниць та різниця між різницями базових елементів значно пом’якшує припущення. Переважно для цього беруться два попередні та два наступні вектори. У результаті отримуємо 39-елементний вектор [4]. Кілька інших можливих варіантів отримання звукових векторів описано в [6].

Звукова модель. Мета звукової моделі – дати метод обчислення правдоподібності будь-якої послідовності звукових векторів при заданій послідовності слів.

У принципі потрібний розподіл імовірностей звукових векторів можна було б знайти, маючи багато зразків кожної послідовності слів та збираючи статистику відповідних послідовностей векторів. Проте це нереально для систем розпізнавання з великим словником.

Замість цього в звуковій моделі послідовності слів розбиваються на базові “будівельні” блоки. Кожен базовий блок подається прихованою моделлю за Марковим (англійська назва – hidden Markov model (HMM)). HMM-модель має формальні вхідний і вихідний стани та ряд породжуючих станів (рис. 2). Вхідний і вихідний стани дозволяють моделям об’єднуватися, щоб утворювати послідовності слів [1, 4, 5].

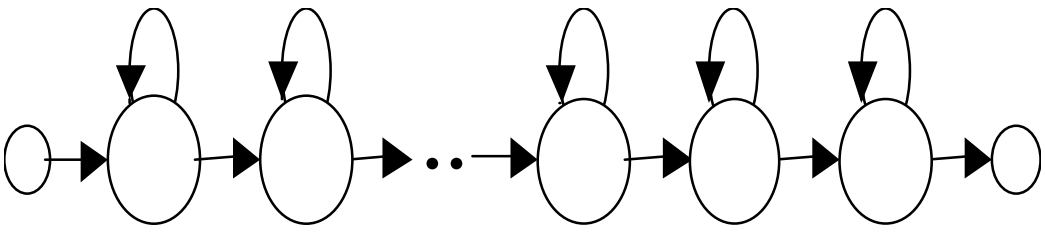

Рис. 2. HMM – модель базових блоків

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Кожен базовий блок являє собою таку фундаментальну одиницю мови як слово, склад чи фонема. Чим простіший опис фундаментальної одиниці, тим менша кількість зв'язаних з нею параметрів.

Ключовим у використанні звукової моделі є зменшення кількості параметрів, які треба оцінити. Це необхідно, бо надто велика кількість оцінюваних параметрів приводить при обмежених навчальних даних до нереальних оцінок. Крім того, зменшення кількості параметрів зменшує обчислювальну складність.

Ця проблема настільки серйозна, що використовуються додаткові еврістичні обмеження [4, 8]:

1. Розподіли ймовірностей появи звукових векторів для різних станів моделей можуть бути зв'язані, тобто користуватися тими самими параметрами. Це корисно лише тоді, коли вони представляють подібні звукові ситуації [8, 9].

2. Коваріаційна матриця розподілів вважається діагональною.

3. Кількість розподілів Гаусса, сума яких моделює розподіл імовірностей для стану, може змінюватися, щоб досягти найкращого балансу між гнучкістю моделювання і складністю.

Перед тим, як HMM може бути застосована, повинні бути визначені її параметри. Цей процес називають навчанням. Він вимагає три елементи.

1. Навчальну базу даних, у якій є мовні записи та відповідні їм тексти.

Для англійської мови існує багато загальнодоступних баз даних мовних зразків (ISOLET, CONNEX,Resource Management Database, Wall Street Journal Database та інші [1-5]). Створення таких баз даних для української мови є завданням на майбутнє.

Як для української, так і для англійської мов завданням є якнайкраще розбиття мовного запису на фонеми відповідно до тексту.

2 Цільову функцію, яка разом з навчальною базою даних може бути використана, щоби виміряти "відповідність" HMM.

Найбільш широко вживаними є три типи цільових функцій: максимальної правдоподібності, максимальної взаємної інформації, міжінтервальної відмінності. Завдання – вибрати одну з них.

3. Процедуру оптимізації, яка може бути використана, щоби максимізувати цільову функцію.

При цьому найчастіше використовується рекурсивний у часі алгоритм Баума-Велча для повторного оцінювання параметрів моделі [1, 4].

Під час навчання процедура оптимізації використовується, щоби знайти вектор параметрів HMM, який має високу відповідність.

Модель мови. Мета моделі мови – дати метод обчислення апріорної ймовірності послідовності слів незалежно від спостереження мовного сигналу. Для цього треба забезпечити механізм оцінки ймовірності певного слова у фразі, якщо знаємо попередні слова.

Простий, але ефективний шлях [4] зробити це – використати N-ки слів, у яких приймається, що дане слово залежить лише від попередніх (N-1)-слів. N-ки слів одночасно мають граматику, смисл і предметну область та зосереджуються на локальних залежностях. Більше того, розподіли ймовірностей для N-ок можна обчислити просто з текстових навчальних даних. Тому не потрібно мати такі точні лінгвістичні правила, як формальна граматика мови.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

У принципі, N-ки можна оцінити простим підрахунком частоти повторюваності слова в навчальних текстах. Як правило, приймають N = 3.

Проблема полягає в тому, що при словнику L слів є L^3 можливих трійок. Навіть для помірного словника в 5000 слів це дуже велика кількість. Тому багато трійок не з'явиться в навчальних даних, а багато інших з'явиться лише раз чи двічі. Внаслідок цього отримані для них оцінки будуть нереальними.

Підхід до вирішення цієї проблеми полягає в тому, що оцінки трійок, які найчастіше з'являються, зменшуються, а отримана залишкова ймовірнісна маса розподіляється між трійками, що рідко зустрічаються [4].

Класифікатор. Ця складова системи зводить воєдино дані від трьох раніше описаних компонент і знаходить найбільш імовірний текст (транскрипцію).

Як правило, усі можливі гіпотези відслідковуються паралельно. Цей підхід спирається на принцип оптимальності Белмана (динамічного програмування) і його часто називають алгоритмом Вітербі [4, 7].

Через складність сучасних систем розпізнавання мови суттєвим завданням є звуження області пошуку найбільш імовірної гіпотези.

Обчислення за алгоритмом Вітербі ведуться послідовно (рекурсивно) в часі. Для обмеження пошуку вводиться поняття активного стану. Оцінка V(t) правдоподібності гіпотези в момент часу t обчислюється, коли вона досяжна з активного стану в момент часу t-1. Активні стани – це такі стани гіпотез, для яких оцінка V(t-1) близька до max V(t-1). Якщо ретельно вибрати поріг, який задає близькість оцінок V(t-1) та max V(t-1),то цей евристичний прийом значно зменшує обсяг обчислень при незначному погіршенні точності розпізнавання.

1. Kapadia S. Discriminative training of hidden Markov models. PhD thesis, Cambridge University, 1998. 2. Hain T., Woodland P.C., Niesler T.R., Whittaker E.W.D. The 1998 HTK system for transcription of conversational telephone speech. Proc. ICASSP'99. – P.57–60. 3. Hain T., Woodland P.C., Evermann G., Povey D. The CU-HTK March 2000 Hub 5E transcription system. Proc. Speech Transcription Workshop. College Park, 2000. 4. Young S. Large vocabulary continuous speech recognition. IEEE Signal Processing Magazine, 13(5), 1996. – P.45-57. 5. Young S. The HTK hidden Markov model toolkit: design and philosophy. Technical Report CUED / F-INFENG/TR152, 1994. 6. Вінцюк Т.К. Аналіз, розпізнавання й інтерпретація мовних сигналів, К., 1987. 7. Young S., Russell N., Thornton J. Token Passing: a simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Department. – 1989. 8. Young S., Odell J., Woodland P.C. Tree-based state tying for high accuracy acoustic modeling. Proc. Human Language Technology Workshop, Plainsboro NJ, Morgan Kaufmann Publishers Inc. 1994. P.307–312. 9. Zhao J., Zhang X., Ganapathiraju A., Deshmukh N., Picone J. Decision tree-based state tying for acoustic modeling. Mississippi State University, Institute for Signal and Information Processing. – 1999.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua