Алгоритм розпізнавання злитого мовлення з надвеликих словників із застосуванням вибірки інформації з баз даних

Валерій Пилипенко

Міжнародний науково-учбовий центр інформаційних технологій та систем,
г. Київ, Україна
valery_pylypenko@mail.ru

Abstract

This paper presents a new two-pass algorithm for Extra Large (more than 1M words) Vocabulary COntinuous Speech recognition based on the Information Retrieval (ELVIRCOS). The principle of this approach is to decompose a recognition process into two passes where the first pass builds the word subset for the second pass recognition. Word graph composition for continuous speech is presented. Experimental results for speech recognition system with vocabulary of about all words (approximately 2 M) are presented.

Вступ

Для розпізнавання мовлення із великих словників для ізольовано (або дискретно) вимовлених слів на початку 90-х років було розроблено декілька систем з достатньо гарними показниками. Потім увага дослідників перенеслася на розпізнавання злитого мовлення, де використовувалися статистичні залежності в порядку слів, що дозволяло передбачити поточне розпізнаване слово на основі кількох попередніх слів. Розроблені статистичні методи прогнозу дозволили значно зменшити кількість альтернатив при розпізнаванні, що у свою чергу дозволило розробляти системи розпізнавання з сумарним обсягом у десятки тисяч слів, хоча на кожному кроці розпізнавання розглядалося декілька сотень альтернатив.

Проте, існує необхідність побудови систем розпізнавання мови з великою кількістю альтернатив і за умови, що немає яких-небудь обмежень на порядок розпізнаваних слів.

Наприклад, при керуванні комп'ютера голосом неможливо передбачити наступне слово на основі декількох попередніх, оскільки це визначається логікою керування комп'ютера, а не властивостями тексту. З іншого боку існує необхідність значного збільшення обсягу словника для того, щоб охопити всі синоніми однієї і тієї ж команди, оскільки користувачу, звичайно, важко запам'ятати тільки один варіант назви команди.

Другий приклад пов'язаний з диктуванням текстів. Використання таких систем, звичайно, обмежено такими текстами, що аналогічні тим текстам, для яких накопичувалися статистики. Крім іншого, додаткове редагування набраного тексту вимагає наявності всіх слів в активному словнику.

Таким чином, існують додатки, де бажано мати словник максимально великого розміру, щоб у майбутньому охопити всі слова даної мови.

Додаткова інформація для обмеження числа альтернатив може бути одержана безпосередньо з мовного сигналу. Для цього пропонується виконати пробне розпізнавання за допомогою фонетичного стенографа. Одержана послідовність фонем формує потік запитів до бази даних для отримання невеликої кількості слів, які могли б входити в словник розпізнавання, що дозволяє значно скоротити кількість альтернатив для розпізнавання.

Наступні розділи описують новий двопрохідний алгоритм. Спочатку представлена базова система для порівняння із запропонованою системою розпізнавання мови. Потім описані два варіанти алгоритму для ізольованих слів та злитого мовлення. Виконані експерименти показують ефективність запропонованих методів.

1. Базова система розпізнавання мовлення

Запропонований метод можна застосувати в будь-якій системі розпізнавання мовлення, де представлені фонеми, і можна сформувати процедуру фонетичного стенографа. У даній роботі як базова система використовується інструментарій HTK [2] на основі прихованих Марківських моделей (Hidden Markov Model - HMM).

1.1. Попередня обробка мовленнєвого сигналу

Мовний сигнал перетвориться в послідовність векторів ознак з інтервалом аналізу 25 мс і кроком аналізу 10 мс. Спочатку мовний сигнал фільтрується фільтром високих частот з характеристикою $P(z) = 1 - 0.97z^{-1}$ та застосовується вікно Хеммінга. Швидке перетворення Фурье переводить часовий сигнал у спектральний вигляд. Спектральні коефіцієнти усереднюються з використанням 26 трикутних вікон, розташованими в мел-шкалі. 12 кепстральних коефіцієнтів обчислюються за допомогою зворотного косинусного перетворення.

Логарифм енергії додається як 13-й коефіцієнт. Ці 13 коефіцієнтів розширюються до 39-мірного вектора параметрів шляхом дописування першої та другої різниць від коефіцієнтів сусідніх за часом. Для обліку впливу каналу застосовується віднімання середнього кепстра.

1.2. Акустична модель

Акустичні моделі відображають характеристики основних одиниць розпізнавання. Для акустичних моделей використовуються приховані Марківські моделі з 64 сумішами Гауссівських функцій щільності імовірності. 47 російських контекстно-незалежних фонем моделюються трьома станами Марківського ланцюга з пропусками.
Словник транскрипцій створюється автоматично з орфографічного словника з використанням множини контекстно-залежних правил.

1.3. Показники базової системи

Акустичні моделі навчалися на вибірці з 12 тис. звукових записів із словника в 2037 слів, вимовлених одним диктором. Розпізнавання проводилося на комп'ютері P-IV 2.4 ГГц.

Для перевірки надійності розпізнавання мови було накопичено 1000 окремо вимовлених слів тим же диктором. Послівна надійність розпізнавання і середній час розпізнавання однієї секунди мови для різних розмірів словника приведені в таблиці 1. Оскільки час розпізнавання лінійно залежить від розміру словника, то для словника в 1987 тис. слів його можна оцінити приблизно в 2300 секунд.

Таблиця 1: Результати розпізнавання окремо вимовлених слів базовою системою

Об'єм словника, тис.	1	15	95
Надійність, %	99.9	97.9	94.7
Час, сек.	1	16	115

Для перевірки надійності розпізнавання злитого мовлення було додатково накопичено 1000 фраз з числами від 0 до 999. Послівна надійність розпізнавання і середній час розпізнавання однієї секунди мови для різних розмірів словника приведені в таблиці 2.

Таблиця 2: Результати розпізнавання злитого мовлення базовою системою

Об'єм словника, тис.	1	15	95
Надійність, %	98.0	96.5	92.6
Час, сек.	2.1	36	205

2. Алгоритм ELVIRS для окремо вимовлених слів

2.1. Архітектура

Архітектура системи розпізнавання ELVIRS (Extra Large Vocabulary Speech recognition based on the Information Retrieval) показана на рис. 1. Такі блоки з базової системи як *обчислення ознак* і *акустичних моделей* використовуються перед першим проходом алгоритму. Також на другому проході використовується звичайне *порівняння образів* в умовах обмеженого словника

Зміни торкаються введення першого проходу алгоритму, де *фонетичний стенограф* використовується для отримання послідовності фонем. Потім процедура *вибірки інформації* створює обмежений словник (підсловник) для другого проходу алгоритму.

2.2. Фонетичний стенограф

Алгоритм фонетичного стенографа[2] створює фонетичну послідовність для мовного сигналу незалежно від словника. Для цього будується деякий автомат породження фонем, який може синтезувати всі можливі моделі мовних сигналів для послідовності фонем. Потім використовується пофонемне розпізнавання для невідомого мовного сигналу.

Використовуються ті ж контекстно-незалежні моделі фонем, що і в базовій системі розпізнавання.

2.3. Процедура отримання підсловника з бази даних

Заздалегідь в процесі навчання із словника транскрипцій створюється індекс від трійок фонем до транскрипцій. Ключем індексу є трійка фонем. Таким чином, таблиця індексу складається з M^3 входжень, де M є число фонем в системі. Кожне входження в таблицю містить список транскрипцій, в які входить трійка фонем ключа входження.

Процес отримання підсловника ілюструється на рис. 2. Вихід фонетичного стенографа ділиться на трійки фонем із зрушенням на одну фонему. Трійка фонем стає запитом до бази даних. Зараз використовується простий запит, коли він в точності співпадає з трійкою фонем. В майбутньому пропонується використовувати відстань *Levensteine* для врахування вставок, видалень та замін в послідовності фонем. Таким чином, послідовність фонем продукує потік запитів до бази даних.

Відповідь на один запит складається із списку транскрипцій, в які дана трійка фонем входить. Цей список копіюється в підсловник для другого проходу алгоритму. Наступний запит з потоку додає нову порцію транскрипцій, при цьому підраховується кількість повторень для того, щоб можна було обчислити ранг слова в підсловнику.

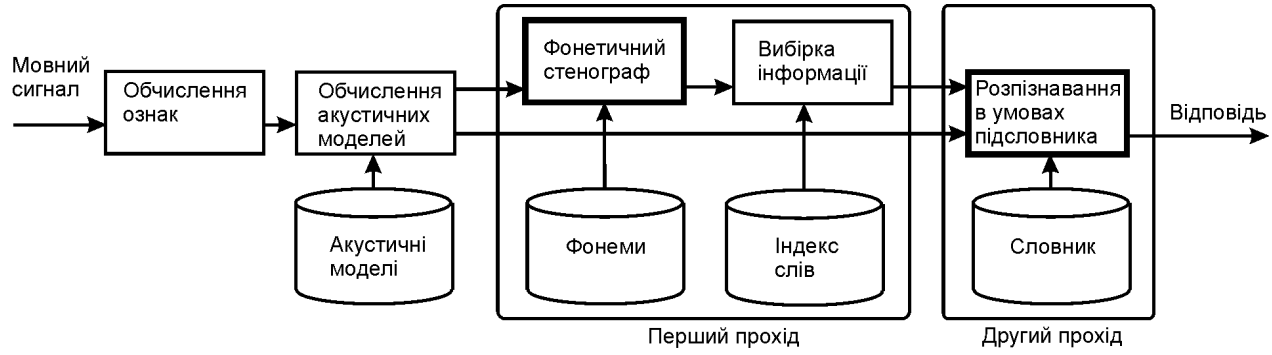

Рис. 1. Архітектура двопрохідної системи розпізнавання мовлення з надвеликих словників

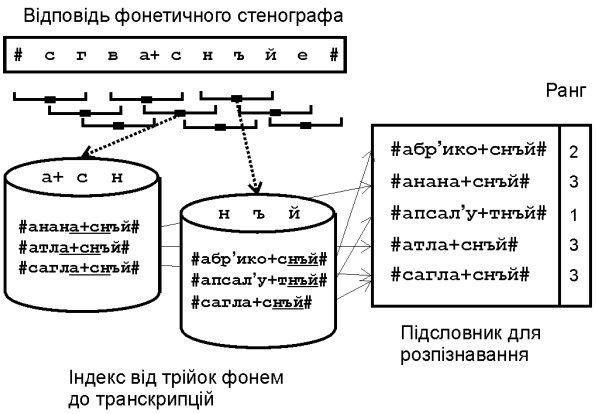

Рис. 2. Процес отримання підсловника

Всі транскрипції в одержаному підсловнику упорядковуються згідно рангу слова (лічильнику повторень). Перші N транскрипцій заносяться в остаточний підсловник для другого проходу алгоритму. Таким чином, підсловник для розпізнавання містить транскрипції з найвищими рангами і число транскрипцій не перевищує фіксованого числа N.

2.4. Алгоритм ELVIRS

Алгоритм ELVIRS складається з двох частин.
Підготовчий етап:

- Підготувати словник для розпізнавання.
- Вибрати множину фонем і створити транскрипції слів із словника за допомогою правил.
- Створити індекс бази даних від трійок фонем до транскрипцій.
- Навчити акустичні моделі по накопичених мовних сигналах.

Етап розпізнавання:

- Застосувати фонетичний стенограф до вхідного сигналу для отримання послідовності фонем.
- Поділити послідовність фонем на трійки фонем із зрушенням в одну фонему.
- Створити запити до БД з трійок фонем
- Одержати списки транскрипцій за допомогою запитів до індексу бази даних.
- Упорядкувати транскрипції до їх рангу.
- Вибрати перші N транскрипцій з найвищими рангами як підсловник для розпізнавання.
- Розпізнати вхідний мовний сигнал в умовах обмеженого підсловника.

3. Інформаційна оцінка імовірності правильного формування підсловника

Відповідь розпізнавання фонетичного стенографа може розглядатися як правильна послідовність фонем, пропущена через канал з шумом. Позначимо у відповіді фонетичного стенографа правильну фонему як 1, а зіпсовану шумом як 0. Нехай імовірність появи 1 в двійковому наборі дорівнює u. Імовірність P знайти в двійковому наборі довжини n підряд k одиниць і більше можна обчислити за допомогою наступного рекурентного виразу:

$$P_n = \begin{cases} 0, & n < k \\ u^k, & n = k \\ P_{n-1} + u^k(1-u)(1-P_{n-k-1}), & n > k \end{cases}$$

У таблиці 3 показана імовірність P знайти в двійкових наборах підряд три і більше 1 при деяких довжинах n та імовірності u. Середня довжина транскрипцій дорівнює приблизно 8 і імовірність правильного знаходження фонеми у відомих реалізацій приблизно дорівнює 85%. При таких значеннях імовірність знайти правильне слово в підсловнику дорівнює 0.953.

Таблиця 3: Імовірність знайти підряд три і більше 1 в двійковому наборі довжини n

n \ u	0.75	0.8	0.85	0.9
6	0.738	0.819	0.890	0.948
7	0.799	0.869	0.926	0.967
8	0.849	0.908	**0.953**	0.982
9	0.887	0.937	0.971	0.991
10	0.915	0.956	0.981	0.995

4. Алгоритм ELVIRCOS для розпізнавання злитого мовлення

4.1. Архітектура

Після отримання списків транскрипцій використовується додаткова процедура *формування графа слів* для злитого мовлення, яка створює мережу слів для другого проходу алгоритму.

4.2. Формування графа слів

Процес створення графа слів показаний на рис. 3. Мережа слів починається з вершини S і закінчується у вершині F. Кожна трійка фонем з відповіді фонетичного стенографа породжує проміжну вершину з номером синхронним до часу появи цієї трійки фонем. З іншого боку кожна трійка фонем стає запитом до індексу бази даних, який повертає список транскрипцій. Транскрипції вставляються між проміжними вершинами так, щоб трійки фонем опинилися в одній колонці по вертикалі.

У випадку, коли відбувається перетин транскрипцій одного слова, породженими різними трійками фонем, тоді ранги цих транскрипцій збільшуються на одиницю. Для кожного моменту часу можна підрахувати число транскрипцій тих, що входять у цей проміжок часу.

Для зменшення складності графа слів використовується обмеження N для кількості слів в кожен момент часу. При цьому віддаляються слова з малими рангами.

Оскільки граф слів формується зліва направо можна проводити його формування у реальному часі із затримкою, яка дорівнює максимальній довжині транскрипції.

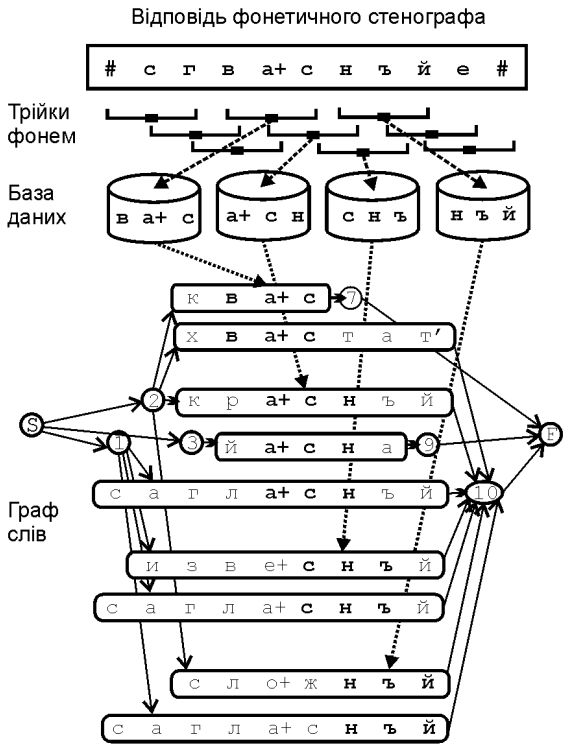

Рис. 3. Формування графа слів для злитого мовлення

4.3. Алгоритм ELVIRCOS

Алгоритм ELVIRCOS складається з двох частин.

Підготовчий етап такої ж, як і в алгоритмі ELVIRS.

Етап розпізнавання:

- Застосувати фонетичний стенограф до вхідного сигналу для отримання послідовності фонем.
- Поділити послідовність фонем на трійки фонем із зрушенням в одну фонему.
- Створити запити до БД з трійок фонем
- Одержати списки транскрипцій за допомогою запитів до індексу бази даних.
- Створити граф слів для злитого мовлення.
- Упорядкувати транскрипції до їх рангу.
- Вибрати перші N транскрипцій з найвищими рангами як підсловник для розпізнавання.
- Розпізнати вхідний мовний сигнал для графа злитої мови в умовах обмеженого підсловника.

5. Експериментальні результати

Для того, щоб ввести перший прохід алгоритму ELVIRCOS в базову систему розпізнавання мови були зроблені необхідні зміни в інструментарії HTK і проведені декілька експериментів.

Для окремо вимовлених слів досліджувався вплив обмеження N на середній час і надійність розпізнавання мовлення для словників різного об'єму, що наведені в таблиці 4. Результати показують корисність обмеження N для словників великих об'ємів, що дозволяє додатково скоротити час розпізнавання при незначному погіршенні надійності.

В цілому одержано значне скорочення часу розпізнавання в сотні разів при відносно невеликому (близько 5%) погіршенні надійності в порівнянні з базовою системою розпізнавання. Погіршення надійності має хороший збіг з оцінкою імовірності правильного формування підсловника.

Таблиця 4: Результати алгоритму ELVIRS

Об'єм словника, тис / Обмеження N на розмір підсловника	15		95		1987	
	Надійн. %	Час, сек.	Надійн. %	Час, сек.	Надійн. %	Час, сек.
50	92.2	1.4	81.0	1.4	69.2	1.6
200	94.6	1.6	87.6	2.1	76.0	1.9
500	**95.5**	**1.9**	**90.1**	**2.5**	80.0	3.3
1000	96.0	2.1	90.7	3.1	82.7	4.4
2000	96.0	4.4	92.0	4.5	**84.8**	**6.8**
5000	96.0	4.6	92.9	8.3	86.4	12.0

Для злитого мовлення були проведені попередні експерименти, в яких розглядався випадок, коли обмеження N співпадало з розміром словника. У таблиці 5 наведені показники надійності для різних розмірів словника.

Таблиця 5: Результати алгоритму ELVIRCOS

Об'єм словника, тис	15	95	1987
Надійність, %	85.3	84.9	83
Час, сек	2.3	9.4	160

Зменшення середнього часу розпізнавання не настільки значне, як у випадку окремо вимовлених слів, оскільки деякі послідовності фонем від фонетичного стенографа породжують графи слів дуже великої розмірності. Сподіваємося, що введення обмеження N істотно зменшить час розпізнавання.

Висновки

Стаття описує новий підхід до розпізнавання мови з великих словників і представляє експериментальну перевірку запропонованих підходів. Слід підкреслити важливість підходів вибірки інформації з баз даних для процесу розпізнавання мови. Показано, що додаткове джерело інформації, одержане з аналізу послідовності фонем від фонетичного стенографа, дозволило значно скоротити простір пошуку в алгоритмі розпізнавання мови. Це дозволило створити системи розпізнавання мови, що охоплюють практично всі слова мови.

Література

[1] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, “The HTK Book”, *Cambridge University Engineering Department,* 2002.

[2] Пилипенко В.В. Використання фонетичного стенографа при розпізнаванні мовлення з великих словників // Тези 12-й міжнародної конференції “Автоматика - 2005”, Харків, 2005, с. 73.