ФОНЕМНА ФІЛЬТРАЦІЯ СИГНАЛІВ МОВЛЕННЯ НА ПІДСТАВІ КЛАСТЕРНОГО АНАЛІЗУ

Стасевич П. А., Тертичний Г. М., Павлов О. І.

Анотація

This paper describes an implementation of the vector quantinizing theory in order to use statistic differences of speech elements and make decision towards a classification (recognition) of observed elements. Time duration and the order of training elements are ignored, while optimality of etalon elements depends on statistic only, and does not depend on temporal rules or other limitation, that have effect in the method of recognition which uses a lot of drawings. There are given results of a carried out experiments.

1. Задача навчання розпізнаванню і супутні задачі, що виникають

З літератури [1] відомо, що при розпізнаванні сигналів мовлення (як слів, що промовляються окремо, так і слів, що зливаються одне з іншим) широко застосовується поелементний спосіб розпізнавання з попереднім *навчанням розпізнаванню за однією реалізацією*. Елементами розпізнавання при цьому вважаються параметри сигналу мовлення, які відповідають стану мовленнєвого тракту людини на кожному елементарному часовому інтервалі. Такі елементи мовлення суттєво залежать від фізичного та емоційного стану диктора і можуть значно різнитися при повторних промовляннях слів навіть з приблизно однаковим темпом.

Поелементний спосіб розпізнавання з попереднім навчанням розпізнаванню за однією реалізацією є спрощеним підходом до розпізнавання мовлення і не забезпечує стійкого розпізнавання при повторному вимовлянні слів без накладання певних обмежень на спосіб їх вимовляння. Тому, для усунення або зменшення таких обмежень та підвищення ймовірності правильного розпізнавання використовується навчання розпізнаванню за багатьма реалізаціями, яке має забезпечувати створення еталонів слів не лише за темпоральними ознаками окремої реалізації, а й з урахуванням статистики елементів і їх темпоральних змін.

Відомий точний розв'язок задачі навчання розпізнаванню за багатьма реалізаціями, який можливо отримати за допомогою методу динамічного програмування [1]. Алгоритм точного розв'язку такої задачі потребує занадто великої кількості повторень рекурентних операцій та обсягів пам'яті для збереження проміжних результатів навіть для навчальної вибірки з 10 реалізацій (при середній тривалості реалізації близько 70 елементарних часових інтервалів кількість повторень сягає 7010, кількість чисел, які потрібно тримати в пам'яті, також сягає 7010). В той же час, навіть однократне повторення таких операцій є занадто великою за обсягами обчислень задачею. Відомий спрощений алгоритм розв'язання задачі навчання розпізнаванню за багатьма реалізаціями [1] не гарантує досягнення глобального оптимуму.

Для зменшення часу, потрібного користувачу для навчання системи розпізнавання, замість повного навчання застосовують часткове навчання, що так само є суттєвим методичним спрощенням і не забезпечує ані стійкого розпізнавання слів скороченої навчальної вибірки, ані певної імовірності правильного розпізнавання решти слів словника.

Для вирішення вказаних проблем навчання розпізнаванню, в тому числі і за багатьма реалізаціями, які є особливо актуальними для великих словників, найбільш ефективним є перехід від навчання розпізнаванню слів до навчання розпізнаванню їх частин, наприклад, фонем. Через те, що кількість фонем обмежена 45-55 фонемами (залежно від мови, діалекту та інших факторів), обсяг фонемної навчальної вибірки абсолютно прийнятний, навіть в разі подання кожної фонеми великою кількістю реалізацій. Це стосується і випадків застосування дифонної або трифонної навчальної вибірки.

Перехід від поелементного розпізнавання слів до пофонемного (дифонного або трифонного) розпізнавання мовлення унеможливлює застосування навчання розпізнаванню за однією реалізацією і висуває задачу продовження пошуку ефективних методів врахування статистичних відхилень окремих реалізацій фонем одна від одної як в часовому просторі, так і в просторі ознак.

2. Постановка задачі

Одним з можливих способів розв'язання такої задачі може бути врахування великої статистики елементів фонем за допомогою їх попереднього векторного квантування, бо, як відомо, векторне (або блочне) квантування є сумісним квантуванням (апроксимацією) блоку параметрів, яке дозволяє ефективно врахувати чотири взаємно пов'язаних характеристики елементів таких векторів: лінійну залежність (кореляцію), нелінійну залежність, форму функції щільності імовірності (ФЩІ) та багатовимірність векторів, в той час як скалярне квантування дозволяє ефективно враховувати лише лінійну залежність та форму ФЩІ [3].

Далі описується підхід щодо застосування теорії векторного квантування для врахування статистичного різноманіття елементів сигналів мовлення з подальшим формування еталонних реалізацій фонем в часовому просторі.

3. Алгоритм розв'язання задачі

Створення статистично оптимальних часових реалізацій еталонних елементів фонем виконується в два етапи.

Перший етап — етап оптимізації еталонних елементів — виконується в просторі ознак за допомогою

кластерного аналізу елементів навчальної вибірки без урахування темпоральних властивостей її елементів.

Другий етап — етап формування темпоральних правил та групування реалізацій — виконується в часовому просторі за допомогою метода динамічного програмування.

Після знаходження еталонних реалізацій фонем будуються так звані фонемні фільтри, задачею яких є розрахунок оцінки схожості фрагменту сигналу мовлення, що спостерігається і відповідних еталонних реалізацій фонем.

На останньому етапі розв'язання задачі приймається рішення щодо класифікації мовленнєвого фрагменту на підставі результатів попереднього аналізу.

4. Створення навчальних вибірок фонограм

При побудові так званих фонемних фільтрів для кожної з можливих фонем найкраще скористатися фонемною базою даних, яка створюється експертами на підставі фонограмної бази достатньо великого розміру та її фонетичного еквіваленту. Така фонемна база має містити велику кількість реалізацій кожної фонеми з достатньою репрезентативністю відповідних фонетичних варіацій, які в подальшому використовують як відповідні навчальні вибірки фонограм.

В проведеному експерименті при створенні фонемного фільтру фонеми „а" через відсутність вказаної фонемної бази даних за попередню навчальну вибірку фонограм були взяті фонограмні записи відкритих двоскладних фрагментів слів, в кожному відкритому складі яких зустрічається така фонема („баба", „база", „вага", „ваза", „вата" та т. п.).

Така навчальна вибірка, крім переважної статистики фонограмних реалізацій звуку „а" в різних літеральних оточеннях, містить ще і фонограмні реалізації інших фонем, статистика яких, хоча і значно менша, але все ж таки буде погіршувати якість побудови кодової книги головної фонеми.

5. Попередня обробка навчальних вибірок фонограм

Попередня обробка навчальних вибірок фонограм для кожної з фонем складається з таких послідовних операцій:

- Сигнал мовлення за допомогою мікрофона та аналого-цифрового перетворювача записується у вигляді фонограми з частотою дискретизації 8000 Гц, що забезпечує якість стандартного телефонного каналу зв'язку і дає змогу зберігати в запису перші три форманти спектрального представлення мовлення. Кодування сигналу мовлення — лінійне імпульсно-кодове (PCM), розрядність — 16 біт.

- Аналіз фонограм проводиться за кадровим принципом, для чого сигнал мовлення послідовно розбивається на кадри тривалістю 10 мс кожен (80 відліків). Розмір кадру аналізу обирається таким, що відповідає тривалості сигналу мовлення 20 мс (160 відліків). Таким чином, кожен кадр аналізу містить попередній та поточний кадри сигналу мовлення тривалістю по 10 мс кожний. Швидкість зміни кадрів аналізу дорівнює 100 кадрів за секунду — кадри аналізу перекривають один одного на 50%, що забезпечує більш повільну зміну результатів аналізу.

- Кожен кадр аналізу, який утворюється з двох незважених кадрів сигналу мовлення (попереднього і поточного), помножується на вагову функцію Хеммінга,

$$W(k) = 0.54 - 0.46 \cdot \cos\left(2\pi \cdot \frac{k}{159}\right)$$

$k = 0,\ldots 159$, що забезпечує зменшення спектральних спотворень, що виникають в разі кадрової обробки сигналів мовлення.

- Зважений кадр аналізу описується одним з можливих наборів параметрів у відповідному просторі ознак, які розраховуються на підставі лінійного прогнозування [2]:

нормовані значення лінійної автокореляційної функції (ACF),

коефіцієнти лінійного прогнозування (LPC),

коефіцієнти відбиття або коефіцієнти драбинного фільтру (RFL),

коефіцієнти логарифмів відношення площин (LAR),

коефіцієнти кепстру (CPS),

коефіцієнти лінійних спектральних проекцій (LSP),

коефіцієнти лінійних спектральних частот (LSF),

коефіцієнти лінійних спектральних проекцій найвищої регресії (PHI),

коефіцієнти лінійних спектральних частот найвищої регресії (FHI).

Кожен з наборів параметрів містить одну й ту саму кількість інформації про спектральну обвідну сигналу мовлення в межах кадру аналізу і може бути перерахований один в інший без втрат. Властивості кожного простору ознак суттєво різняться, що по-різному впливає на результати розв'язання специфічних задач обробки мовлення в кожному окремому випадку і залишається об'єктом численних досліджень.

Ще одним з вдалих наборів параметрів є коефіцієнти згладженого енергетичного спектру, значення яких можна знайти за допомогою перетворення Фур'є, або розрахувати через відому частотну характеристику мовленнєвого тракту.

Оцінка тимчасової частотної характеристики мовного тракту базується на методі лінійного прогнозування [2], основне рівняння якого таке:

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{m1} \\ C_{12} & C_{22} & \cdots & C_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1m} & C_{2m} & \cdots & C_{mm} \end{bmatrix}^{-1} \times \begin{bmatrix} -C_{01} \\ -C_{02} \\ \vdots \\ -C_{0m} \end{bmatrix},$$

де α_i — коефіцієнти лінійного прогнозування, $C_{i,j} = \sum_{n=n_1}^{n_2} S_{n-i} \cdot S_{n-j}$ — коефіцієнти кореляції, а S_i — відліки сигналу мовлення, отримані в результаті дискретизації, n_1, n_2 — межі інтервалу мінімізації середньоквадратичної похибки лінійного прогнозування.

Частіше за все обирається порядок прогнозування m = 10, що забезпечує ефективне представлення спектральної обвідної кадру сигналу мовлення з п'ятьма або меншою кількістю формантних піків.

Перехід від коефіцієнтів лінійного прогнозування до частотної характеристики мовленого тракту відбувається за формулою

$$H(j\omega) = \frac{1}{1 - \sum_{i=1}^{m} (-\alpha_i) \cdot \exp\left(- j \cdot \frac{\omega}{\omega_s} \cdot i 2\pi \right)},$$

де ω_s – частота дискретизації.

Для оцінки коефіцієнтів згладженого енергетичного спектру розраховують значення частотної характеристики мовленого тракту на певних частотах, які обирають за логарифмічним законом так, щоб максимально передати поведінку перших трьох – п'яти формантних піків.

6. Навчальні вибірки векторів ознак

В результаті попередньої обробки навчальних вибірок фонограм створюються навчальні вибірки векторів ознак для кожної фонеми, кожний вектор яких є набором з 10 параметрів обраного простору.

Можливе включення у вектор параметрів додаткових параметрів, таких як значення частоти основного тону, потужність, ознака тон/шум, та інші, які оцінюються додатково в разі необхідності.

Навчальні вибірки векторів ознак містять послідовність векторів, що описують послідовну зміну параметрів мовлення в обраному просторі та додаткових параметрів, якщо такі є.

7. Створення фонемних кодових книг

Кожна фонемна кодова книга будується на підставі відповідної навчальної вибірки векторів ознак.

Для створення оптимальних кодових книг застосовується відомий алгоритм К-середніх [3]. Через те, що розміри кодових книг попередньо не визначені і з'ясовуються за результатами їх синтезу, використовується split-алгоритм поступового збільшення кількості кластерів у векторному просторі, подібний до LBG-алгоритму, запропонованому Ліндо, Бузо, Греєм [3], який описується наступною послідовністю дій:

крок 1: визначення центрального вектору (центроїду) за принципом середнього арифметичного від координат усіх векторів навчальної вибірки;

крок 2: розщеплення центроїда на два нових кодових вектори, які є першим наближенням до двох нових центроїдів;

крок 3: класифікація кожного вектору навчальної вибірки, щодо найближчого кодового вектору; кожний кодовий вектор набуває своїх найближчих сусідів, які утворюють поточний кластер;

крок 4: корегування положення центроїду: переміщення кожного кодового вектору в центр свого кластеру; після цього знову уточнюються найближчі сусіди для кожного кодового вектору і корегуються координати центроїдів, — повторюються кроки 3 та 4; повторення кроків 3 та 4 слід проводити до тих пір, поки центроїди не перестануть істотно рухатись у m - вимірному просторі, а зміна відносної сумарної квадратичної похибки квантування не стане меншою наперед заданого значення;

крок 5: аналогічний кроку 2 – тільки тепер обирається центроїд з найбільшою сумарною квадратичною похибкою, який розщеплюється на два кодових вектори, після чого повторюються кроки 3 і 4.

Такі цикли виконуються необхідну кількість разів, доки не буде отримана кодова книга з задовільними властивостями.

В результаті створення окремих фонемних кодових книг весь векторний простір кожної навчальної вибірки векторів ознак буде розбито на певну кількість кластерів K, будуть знайдені кількість векторів, що утворюють кожний кластер L_k, значення кодових векторів або центроїдів C_k — середніх векторів кожного з кластерів. Імовірності появи кожного центроїду при кодуванні навчальної вибірки розраховуються за формулою

$$p_k = \frac{L_k}{\sum_{i=1}^{K} L_i}.$$

Ентропія всієї кодової книги:

$$H = -\sum_{i=1}^{K} p_k \log_2 p_k.$$

Сумарні квадратичні похибки кодування (апроксимації) векторів кожного кластеру відповідним центроїдом:

$$D_k(x_{k,i}, C_k) = \sum_{i=1}^{L_k} \left(x_{k,i} - C_k \right)^2,$$

де $x_{k,i}$ — вектора навчальної вибірки, що класифікуються як найближчі сусіди до центроїду C_k: $x_{k,i} \in C_k$.

Дисперсії кожного кластеру:

$$\sigma_k^{\,2} = \frac{1}{L_k} \sum_{i=1}^{L_k} \left(x_{k,i} - C_k \right)^2,$$

де $x_{k,i} \in C_k$, інтервал довіри для кожного кластеру, $\Delta_k = \sigma \cdot t_{s,p}$, де $t_{s,p}$ — коефіцієнт Ст'юдента-Фішера

для певного розміру вибірки та заданої імовірності довіри.

8. Додаткова фільтрація навчальних вибірок

Дослідження побудови кодових книг параметрів сигналів мовлення свідчать про те, що незважаючи на критерій оптимальності, який використовується — сумарно квадратичний (мінімізується сумарна квадратична похибка всієї кодової книги) або мінімаксний (мінімізується найбільша сумарна квадратична похибка будь-якого кластеру), — наповненість кластерів векторами навчальної вибірки є різною, а сумарна квадратична похибка кластерів — майже однакова. Це пояснюється тим, що розміри кластерів сильно різняться один від одного: великі за геометричними розмірами кластери містять малу кількість сильно розпорошених навчальних векторів і дають таку саму сумарну квадратичну похибку як і малі за геометричними розмірами кластери, які містять велику кількість сильно згущених навчальних векторів. Саме це і забезпечує головний принцип поведінки векторного квантування, як наближення до ентропійного кодування: кодова книга буде давати з найбільшою імовірністю досить малі похибки кодування, а значні похибки кодування будуть мати значно меншу імовірність [3].

Якщо навчальна вибірка складається з векторів, що мають суттєво різну імовірність — досить значну для векторів ознак головної фонеми і досить малу для векторів ознак фонем-оточень, — розміри кластерів для таких векторів будуть теж суттєво різнитися, як геометрично, так і за наповненістю. Це дає підстави провести зменшення розмірів кодової книги шляхом викидання кластерів з малою кількістю векторів і великою дисперсією, що можна розглядати як додаткову фільтрацію навчальної вибірки. Таким чином можна зменшувати залежність кодових книг від векторів ознак фонем-оточень.

9. Створення фонемної структури та формулювання темпоральних правил

Для створення фонемної структури спочатку відтворюють вихідні реалізації фонемних фільтрів. Для цього кожна часова реалізація навчальної вибірки векторів ознак квантується за допомогою відповідної векторної кодової книги і перетворюється у часову реалізацію оптимальних еталонних елементів — кодових векторів (центроїдів), або їх індексів. Послідовність зміни центроїдів або їх індексів утворює можливі траєкторії відповідної фонеми та визначає правила темпоральної транскрипції (рис. 1).

На рис. 1 наведено приклад структури фонеми та правила зміни центроїдів C_i. Так, траєкторія фонеми може починатися з центроїду C_1 або C_2 і закінчуватися

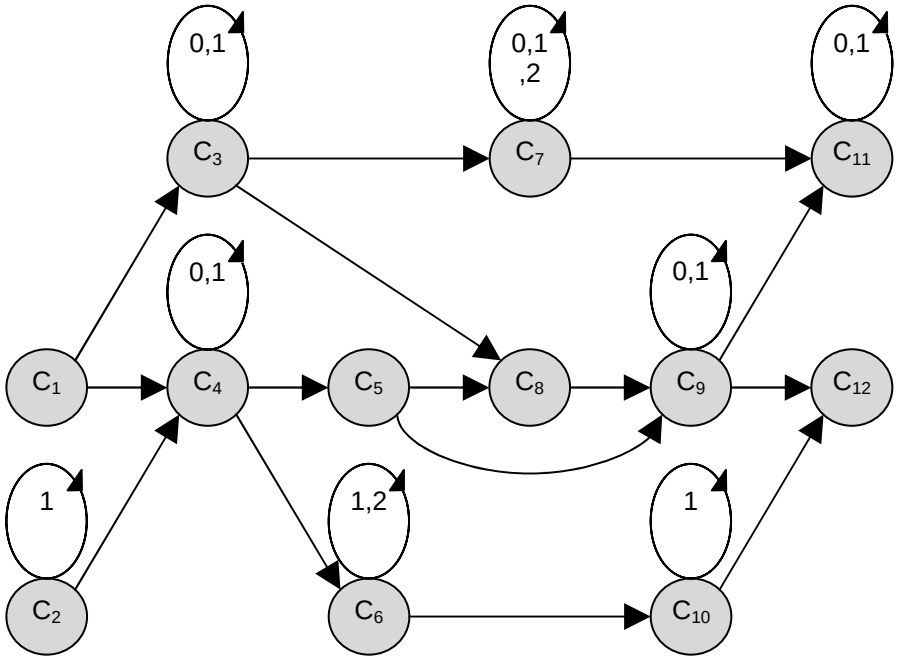

Рис. 1 Приклад структури фонеми

центроїдом C_{11} або C_{12}. Деякі центроїди можуть повторюватися декілька разів, наприклад центроїд C_3 може повторюватися один раз або не повторюватися жодного разу. Деякі центроїди обов'язково повторюються декілька разів, наприклад центроїд C_{10} обов'язково повторюється один раз.

Таблиця 1

Номер траєкторії	Послідовність еталонних елементів та їх темпоральна транскрипція					
1	{C_1;1}	{C_3;1,2}	{C_7;1,2,3}	{C_{11};1,2}		
2	{C_1;1}	{C_3;1,2}	{C_8;1}	{C_9;1,2}	{C_{11};1,2}	
3	{C_1;1}	{C_3;1,2}	{C_8;1}	{C_9;1,2}	{C_{12};1}	
4	{C_1;1}	{C_4;1,2}	{C_5;1}	{C_8;1}	{C_9;1,2}	{C_{11};1,2}
5	{C_1;1}	{C_4;1,2}	{C_5;1}	{C_8;1}	{C_9;1,2}	{C_{12};1}
6	{C_1;1}	{C_4;1,2}	{C_5;1}	{C_9;1,2}	{C_{11};1,2}	
7	{C_1;1}	{C_4;1,2}	{C_5;1}	{C_9;1,2}	{C_{12};1}	
8	{C_1;1}	{C_4;1,2}	{C_6;2,3}	{C_{10};2}	{C_{12};1}	
9	{C_2;2}	{C_4;1,2}	{C_5;1}	{C_8;1}	{C_9;1,2}	{C_{11};1,2}
10	{C_2;2}	{C_4;1,2}	{C_5;1}	{C_8;1}	{C_9;1,2}	{C_{12};1}
11	{C_2;2}	{C_4;1,2}	{C_5;1}	{C_9;1,2}	{C_{11};1,2}	
12	{C_2;2}	{C_4;1,2}	{C_5;1}	{C_9;1,2}	{C_{12};1}	
13	{C_2;2}	{C_4;1,2}	{C_6;2,3}	{C_{10};2}	{C_{12};1}	

З наведеної структури фонеми можна утворити певну кількість еталонних траєкторій такої фонеми з відповідними темпоральними правилами (табл. 1).

10. Результати експериментальних досліджень

При проведенні експерименту було визначено кодову книгу, що відповідає фонемі “а”, тобто програма для тесту мала відрізнити фонему “а” від інших. Сигнал мовлення, що підлягав розпізнаванню, складався з послідовно вимовлених фонем “а”, “о”, “у” та “і” (рис. 2).

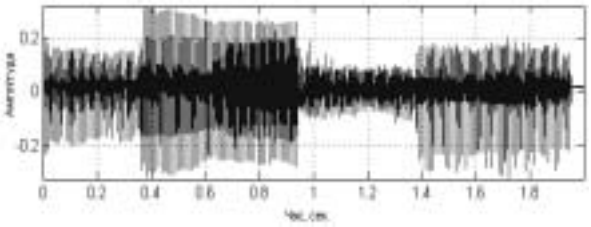

Рис. 2 Амплітудно-часова залежність сигналу мовлення

Програма розпізнавання на основі статистичних даних приймала рішення – чи відповідає вхідний вектор фонемі “а”, і, якщо це так, то на графіку істинності мав з’явитись рівень “1”. З рис. 3 видно, що більш як половина векторів, що відповідали фонемі “а” (перші 0,38 секунди) були розпізнані вірно. Остаточний результат, що можна використовувати в машинному розпізнаванні, отримується після згладжування графіка істинності (Рис. 4).

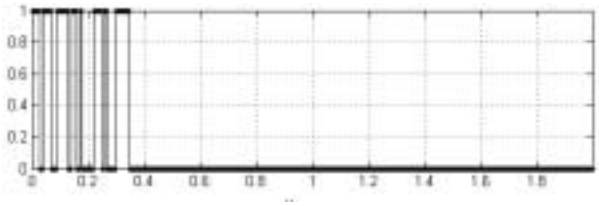

Рис. 3 Графік істинності

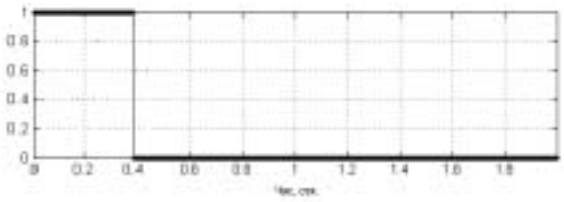

Рис. 4 Графік істинності після згладжування

11. Висновки

Розглянута методика застосування ряду перетворень у комбінації з векторним квантуванням дає можливість враховувати статистичне різноманіття елементів сигналів мовлення. Принципова відмінність методу від тих, що використовувалися раніше полягає в тому, що часова тривалість і послідовність зміни елементів навчальної вибірки не враховувались. Таким чином оптимальність елементів еталонів, що утворюються, залежить лише від поданої статистики, а не від темпоральних правил чи інших припущень та обмежень. Наведені результати експериментів засвідчують практичність застосування.

12. Посилання

[1] Т. К. Винцюк. *Анализ, распознавание и интерпретация речевых сигналов*. Киев: Наукова думка, 1987.

[2] Дж. Д. Маркел, А. Х. Грэй. *Линейное предсказание речи*. Перевод с английского под. ред. Ю. Н. Прохорова и В. С. Звёздина. М.: Связь, 1980.

[3] Y. Linde, A. Buzo, and R. M. Gray, *An algorithm for vector quantizer design*, IEEE Trans. Communications., Vol, Com-28, Jan. 1980.