Моделі навчання і розпізнавання в інформаційній технології розпізнавання мовлення з великих словників

Ольга Савенкова, Олег Карпов

Дніпропетровський національний університет
2sol@ukr.net

Abstract

New information technology for the speech recognition system by syllabic parameters trajectory synthesis is considered. The technology based on two models: learning and recognition. Functional diagrams of the models are presented.

1. Вступ

Створення систем розпізнавання мовлення (СРМ) з великим словником потребує: визначення структури об'єктів розпізнавання та операцій, необхідних для інтелектуального розв'язання проблеми; розроблення швидкозбіжних стратегій для ефективного пошуку потенційних розв'язків, які можуть бути згенеровані цими структурами і операціями з урахуванням додаткової інформації (евристик) про досліджувану проблемну область [1-10]. Відомі технології побудови таких СРМ використовують підхід, який не враховує залежності між параметрами суміжних сегментів, тобто в аналізованому мовленнєвому сигналі (МС) послідовність сегментів розглядається як сукупність незалежних подій, яким відповідають незалежні мовні одиниці (МО) [3-7]. Тому існує необхідність створення інформаційної технології (ІТ), що оперує з такими МО, які можуть використовуватися для композиції слів або речень, задовольняють вимозі максимальної повноти покриття множини слів і враховують залежності між сегментами. Такі властивості мають склади [8, 10, 11].

Модель складового синтезу, яку покладемо в основу пропонованої інформаційної технології (ІТ), формулюється таким чином [10, 13].

Нехай заданий словник $\{SL_k\}$, що містить N МО. Для кожної МО обчислені послідовності параметрів або траєкторії параметрів (ТП) $\{Y_k\}$. ТП для МО SL_k сегментована на n_k ($k = 1 \div N$) сегментів-фонем. Нехай є невідомий МС, для якого обчислені послідовності параметрів X та сегментовані на m_p сегментів-фонем, які можна об'єднати в M сполучень SL_p^x з двох, трьох або чотирьох сегментів ($p = 1 \div M$).

Для розпізнавання МС необхідно знайти композицію з ТП Y_k елементів словника, яка найкраще відповідає послідовності параметрів невідомої реалізації X.

Розробка ІТ для СРМ з великих словників на основі алгоритмів реалізації складового синтезу вимагає розв'язання наступних задач.

- Створення оптимальних словників МО, а саме: вибір способу економного зберігання параметрів МО словника.
- Розробка ефективних алгоритмів розпізнавання на основі складового синтезу послідовності параметрів для невідомого МС.

Відповідно до методології та концепцій створення образного комп'ютера [2], ієрархічних принципів побудови СРМ [10], результатами експериментальних досліджень [12, 13] і з урахуванням рекомендацій до проектування та розробки інформаційної технології, в основу пропонованої ІТ покладемо дві моделі: модель навчання і модель розпізнавання.

2. Модель навчання

Схема моделі навчання пропонованої інформаційної технології зображена на рис. 1 і складається з наступних процедур.

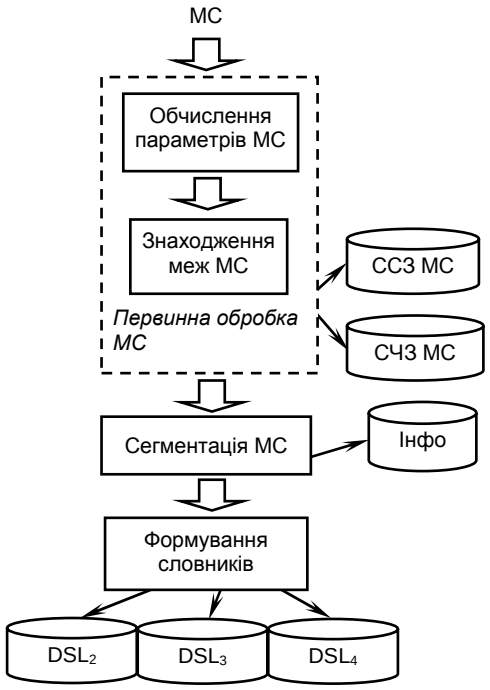


Рисунок 1: Функціонально-структурна схема моделі навчання СРМ

1. Первинна обробка МС.

Первинна обробка МС, який надходить з мікрофона або wav-файлу, полягає в обчисленні інформативних параметрів МС і формуванні траєкторії параметрів $X = \{x_1, x_2, \ldots, x_i, \ldots, x_p\}$ (p – кількість інтервалів аналізу МС; $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in}\}$ – вектор ознак

вимірністю n) і видаленні пауз на початку та кінці мовленнєвого висловлення.

Для подання ТП X досліджуваного МС у даній інформаційній технології обрані [10, 12]:

- спектрально-часове зображення (СЧЗ) $XA(\omega, t)$;
- спектрально-смугове зображення (ССЗ) $XE(l, t)$.

СЧЗ і ССЗ обрані з наступних міркувань [12]: збільшення швидкодії розпізнавання у випадку використання ССЗ у середньому в 9 разів приводить до збільшення похибки розпізнавання на 1,6 % у порівнянні з використанням СЧЗ. Тобто за надійністю розпізнавання обидва зображення ТП практично порівнянні і можуть використовуватися як ієрархія розпізнавання: ССЗ для попереднього розпізнавання та виділення списку кандидатів на розпізнавання, а потім СЧЗ для прийняття остаточного рішення.

2. Автоматична сегментація МС.

Сегментація МС виконується за методом верифікації часової послідовності параметрів [10].

3. Формування словників для мовленнєвих одиниць (МО).

Для формування словника МО виконується автоматичний поділ сегментованої послідовності параметрів МС на акустичні склади-еталони (дво-, три-, чотирисегментні склади, яким відповідають дво- та трисимвольні МО SL_{2k}, $\mathrm{SL}_{3k} \subset \mathrm{SL}_k$, де $k = 1..N_{SL}$) з маркуванням лінгвістичної інформації (назва МО, транскрипція). Траєкторії параметрів МО { YSL^{2k} }, { YSL^{3k} }, { YSL^{4k} } і допоміжна інформація розподіляються в словники відповідно до формату зберігання даних. Інформація про МО у словнику DSL_m ($m = 2, 3, 4$) зберігається у наступному вигляді:

<*Номер МО*>_<*Ім'я МО*>_<*Транскрипція МО*>_
<*Кількість часових відліків*>_<*Кількість сегментів*>_
<*Адреси меж сегментів*>_
<*Групова належність сегментів*>.

Для створення СРМ, що працює в реальному часі, виникає необхідність побудови оптимальних словників МО з погляду об'єму пам'яті, яка використовується для збереження словника, та швидкодії обробки в процесі розпізнавання. Отже, необхідно обрати для кожного зображення ТП модель опису у класі аналітичних функцій, відповідно до якої можна відновити вихідну ТП із мінімальною похибкою. Для цього використовуються наступні методи.

А. Сплайн-опис ССЗ [10]. Для сегментованої ТП $YE(l, t)$ МО словника обчислюються параметри моделі сплайн-опису $a^l_{k,i}$, $b^l_{k,i}$, $c^l_{k,i}$, $d^l_{k,i}$ у кожній частотній смузі l ($l = 1..9$)

$$\tilde{YE}^l = \begin{cases} \tilde{YE}_{1\,i}, & s_0 \le i \le s_1, \\ \dots \\ \tilde{YE}_{k\,i}, & s_{k-1} \le i \le s_k, \\ \dots \\ \tilde{YE}_{N_{SG}\,i}, & s_{N_{SG}-1} \le i \le s_{N_{SG}}, \end{cases} \qquad (1)$$

де для кожного k – го сегмента

$$\tilde{YE}^l_{k,i} = a^l_{k,i} \cdot (\tilde{t}_i)^3 + b^l_{k,i} \cdot (\tilde{t}_i)^2 + c^l_{k,i} \cdot \tilde{t}_i + d^l_{k,i}, \qquad (2)$$

$k = 1..N_{SG}$ (N_{SG} – кількість сегментів); $s_0, s_1, \dots, s_{N_{SG}}$ – межі сегментації; $\tilde{t}_i = t_i - s_{k-1}$.

Б. Опис СЧЗ у класі дзвіноподібних функцій. Для побудови аналітичного опису СЧЗ $YS^*(\omega, t)$ елементів словника використовується функція модифікований локон Аньєзі [14]. За алгоритмом, який запропонований у роботі [14], для елементів словника ТП $YA(\omega, t)$ яких визначені в деякій частотно-часовій області Ω: $[\omega_0,\ \omega_M] \times [t_0,\ t_N]$, обчислюються параметри дзвіноподібних функцій { $Zt(t_l)$ }, { $Z\omega(\omega_k)$ }. Аналітичний опис $YA^*(\omega_k, t_l)$ СЧЗ елемента словника в деякій точці області Ω обчислюється як суперпозиція L добутків дзвіноподібних функцій $Zt_{(i)}(t_l)$, $Z\omega_{(i)}(\omega_k)$ (ω_k, t_l – дискретно задані частота та час, $k = 1..M$, $l = 1..N$, $i = 1..L$), а саме:

$$YA^*(\omega_k, t_l) = \sum_{i=1}^{L} Z\omega_{(i)}(\omega_k) \cdot Zt_{(i)}(t_l), \qquad (3)$$

$$Z\omega_{(i)}(\omega_k) = \frac{b^3_{(i)}}{d^2_{(i)} + (\omega_k - \Omega_{(i)})^2}, \; Zt_{(i)}(t_l) = \frac{a^3_{(i)}}{c^2_{(i)} + (t_l - T_{(i)})^2}.$$

Таким чином, структура оптимального словника МО SL_k складається з: лінгвістичної інформації про МО; допоміжної інформації про ТП (кількість часових відліків, кількість сегментів, межі сегментів, групова належність сегментів Т – "Тон"; Ш – "Шум"; П – "Пауза"); ТП елементів словника у вигляді: параметрів моделей сплайн-опису ССЗ, параметрів моделей опису СЧЗ у класі дзвіноподібних функцій.

Найбільш значимі характеристики сегментно-складового подання мовленнєвої послідовності використовуються для аналізу елементів словника і обчислення евристичних оцінок.

- Вкладеність МО $\mathrm{SL}_{2k} \in \mathrm{SL}_{3k}$, $\mathrm{SL}_{2k} \in \mathrm{SL}_{4k}$: розраховується значення оцінок h_v, $0 \le h_v(n) \le 1$;
- Наявність певної структури групових ознак сегментів: розраховується значення оцінки $h_g(n)$, $0 \le h_g(n) \le 1$; значення оцінки відстані $h_d(n)$ між ТП складів, що містять різні сполучення групових ознак.

На основі розрахованих евристичних оцінок формується структура групування елементів у словниках, завдяки чому скорочується час на пошук найбільш відповідного елемента при розпізнаванні.

3. Модель розпізнавання

Функціонально-структурна схема моделі розпізнавання даної ІТ зображена на рис. 2. Розглянемо основні процедури моделі розпізнавання.

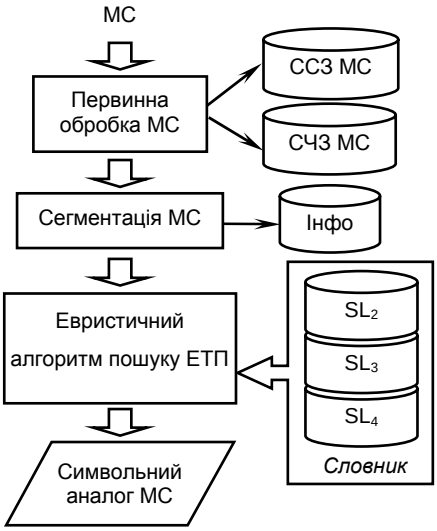


Рисунок 2: Функціонально-структурна схема моделі розпізнавання

До поданої реалізації МС застосовуються процедури первинної обробки та сегментації.

На наступному етапі до сегментованої ТП застосовується евристичний алгоритм сегментно-складового синтезу, згідно з яким виконуються наступні кроки.

1. Процедура напрямленого пошуку кандидатів-композицій для еталонної ТП (ЕТП), які сформовані з ТП складів-еталонів, за допомогою алгоритмів пошуку у ширину та у глибину із використанням евристичних оцінок [10, 13].

У процесі напрямленого пошуку для синтезу ЕТП використовуються подання ТП у вигляді ССЗ, які відновлені із використанням параметрів відповідних елементів із словників DSL_m ($m = 2, 3, 4$) згідно моделі сплайн-опису (1), (2). Сплайн-синтез ЕТП XE^* виконується відповідно до наступної моделі конкатенації ТП складів-еталонів

$$XE^*(l,t) = \begin{cases} \tilde{Y}_i^1, & N'_0 \le i \le N'_1, \\ \dots \\ \tilde{Y}_i^k, & N'_{k-1}+1 \le i \le N'_k, \\ \dots \\ \tilde{Y}_i^M, & N'_{R-1}+1 \le i \le N'_R, \end{cases} \qquad (4)$$

де $k = 1..R$; R – кількість складів-еталонів $\tilde{Y}_i^k$ в ЕТП XE; $\tilde{Y}_i^k$ – ТП деякої МО словника, відновлена відповідно до моделі (1), (2); N_k – кількість часових відліків $k-$ї ТП, $t_1 \in [1, N_1]$, …, $t_k \in [1, N_k]$, …, $t_R \in [1, N_R]$. Межі складів усередині поточної комбінації ТП для ЕТП визначаються таким чином: $N'_0 = 1$; $N'_1 = N_1$; $N'_2 = N_1 + N_2$; …; $N'_R = N_1 + N_2 + \dots + N_R$.

2. Процедура пошуку розв'язку серед обраних кандидатів-композицій для генерування ЕТП. У процесі остаточного прийняття рішення про розпізнавання за критерієм найкращої відповідності використовується подання ТП у вигляді СЧЗ. Генерування ЕТП здійснюється за наступним алгоритмом.

ТП $j-$ї МО визначена в прямокутній частотно-часовій області D^j: $[\omega_0, \omega_M] \times [t_0, t_{Nj}]$ і відновлюється у вигляді суперпозиції добутків дзвіноподібних функцій $Zt_{(i)}(t_l)$, $Z\omega_{(i)}(\omega_k)$ ($i = 1..L_j$, $k = 1..M$, $l = 1..Nj$).

Для поточної композиції ЕТП $XA^*(\omega, t)$, яка містить R МО словника, частотно-часова область визначення з урахуванням об'єднання частотно-часових діапазонів кожної МО – D: $[\omega_0, \omega_M] \times [t_0, t_N]$, де $t_N = t_{N1} + t_{N2} + \dots + t_{NR}$.

Опис СЧЗ для ЕТП $XA^*(\omega, t)$ у точці (ω_k, t_l) області визначення D обчислюється як суперпозиція R гладких функцій $YA_m^*(\omega, t)$ ($m = 1..R$), таким чином

$$XA^*(\omega_k, t_l) = \sum_{m=1}^{R} YA_m^*(\omega_k, t_l), \qquad (5)$$

де $YA_m^*(\omega_k, t_l) = \sum_{i=1}^{Lm} Z\omega_{(i)}(\omega_k) \cdot Zt_{(i)}(t_l)$, R – кількість складів в ЕТП $XA^*(\omega, t)$, L_m – кількість параметрів дзвіноподібних функцій $Zt_{(i)}(t_l)$, $Z\omega_{(i)}(\omega_k)$ ($i = 1..L_m$) для відповідного $m-$го складу комбінації ЕТП $XA^*(\omega, t)$.

3. Композиція символьного аналога МС. Після завершення роботи алгоритму пошуку ЕТП, яка найкраще відповідає послідовності параметрів невідомого МС, виконується композиція її символьного аналога

$$W = f_{compose}(\mathrm{SL}_1, \mathrm{SL}_2, \dots, \mathrm{SL}_i, \dots, \mathrm{SL}_R). \qquad (6)$$

4. Програмна реалізація

Подана у статті інформаційна технологія була реалізована у вигляді комп'ютерної системи розпізнавання мовлення *SPeach* за допомого засобів середовища Borland Delphi 5. Алгоритми та моделі даної ІТ (див. рис. 1, 2) реалізовані в модулях, які перераховані в таблиці 1.

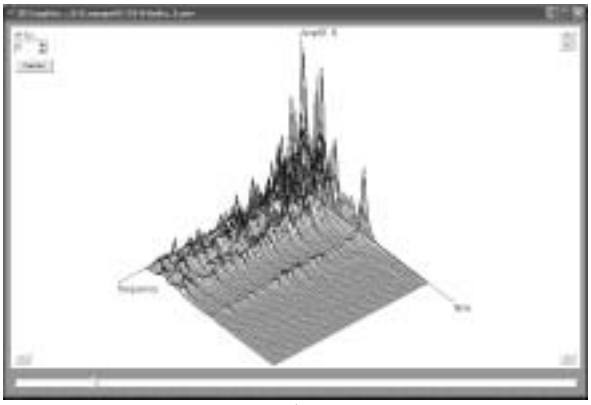
Таблиця 1: Перелік основних модулів системи SPeach

Назва модуля	Призначення
WAVReadWrite	введення МС
SpectrAnalis	обчислення параметрів МС
Segmentation	сегментація МС, визначення типів сегментів
FirmirVocab	формування словників МО
Approx_Spline	побудова сплайн-опису ССЗ
Approx_Anjezi	побудова опису СЧЗ у класі дзвіноподібних функцій
Evristics	евристичний алгоритм
Draw3dGraph	графіка

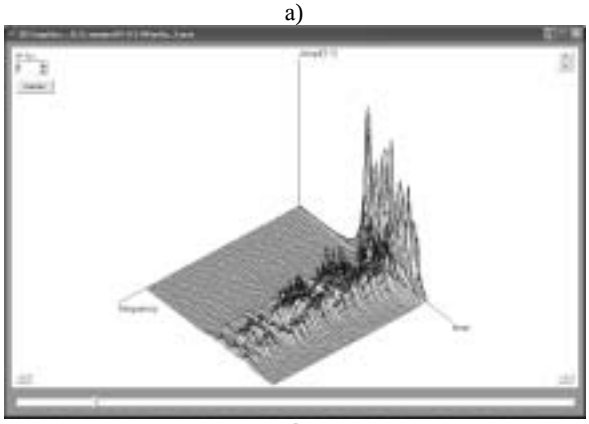
Для тестування розробленої системи був проаналізований словник з 5000 слів. Для цієї множини слів знайдено: 3377 трисимвольних, 555 двосимвольних МО. На етапі формування словників DSL_m ($m = 2, 3, 4$)

ТП цих МО згруповані таким чином: двосимвольні (252 двосегментних, 296 трисегментних, 69 чотирисегментних); трисимвольні (30 двосегментних, 1351 трисегментних, 1587 чотирисегментних). У кожній такій групі ТП проаналізовані на належність сегментів до певного типу Т-Ш-П і обчислені значення евристичних оцінок, які дозволяють скоротити кількість розглядаємих елементів словника.

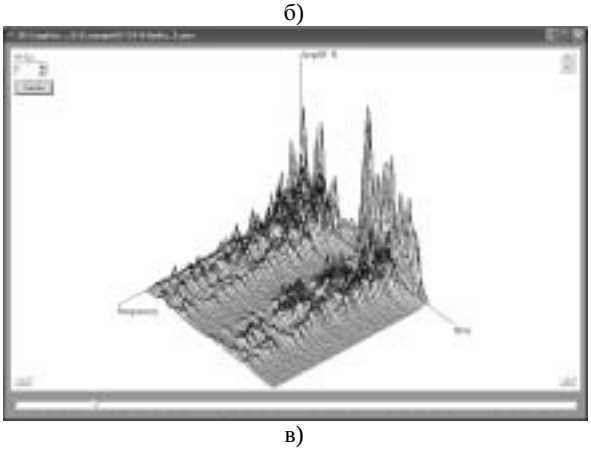
На рис. 3 а), б), в) наведений приклад синтезу ЕТП для поданої реалізації МС "адін" згідно моделі розпізнавання.



а)



б)



в)

Рисунок 3: Опис СЧЗ у класі дзвіноподібних функцій: а) склад "ад"; б) склад "ін"; в) конкатенація ТП складів "ад" и "ін" в ЕТП

Результати тестування розробленої системи *Speech* підтверджують ефективність запропонованої інформаційної технології.

5. Висновки

В даній статті запропонована нова інформаційна технологія для побудови інтелектуальної системи розпізнавання мовлення, яка забезпечує комплексний підхід, що враховує взаємозв'язки між ієрархією подання інформації про мовленнєвий сигнал. В наступному передбачається удосконалення евристичного алгоритму пошуку з метою підвищення швидкодії розпізнавання.

6. Література

[1] Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – К., Наукова думка, 1987. – 264 с.

[2] Вінцюк Т.К. Образний комп'ютер // Сучасні проблеми в комп'ютерних науках. Зб. наук. праць, Вид-во Нац. ун-ту «Львівська політехніка», Львів, 2000, с. 5-14.

[3] Ronzhin A.L., Yusupov R.M, Li I.V., Leontieva A.B. Survey of Russian Recognition Systems. // In Proc. SPECOM'2006, St. Petersburg, 2006, pp. 54-60.

[4] Ронжин А.Л., Карпов А.А., Ли И.В. Система автоматического распознавания русской речи SIRIUS // Искусственный интеллект. – 2005. – № 3. – С. 590-601.

[5] Meisel W.S., Anikst M.T., Pirzadeh S.S., Schumacher J.E., Soares M.C., Trawick D.J. The SSI large-vocabulary speaker-inependent continuous speech recognition system // In Proc. ICASSP'91 Int. Conf. Acoust., Speech and Signal Process., Toronto, 1991, pp. 337-340.

[6] Пилипенко В.В. Распознавание дискретной и слитной речи из сверхбольших словарей на основе выборки информации из баз данных // Искусственный интеллект. – 2006. - №3. – С. 548-558.

[7] Tatarnikova M., Tampel I., Oparin I., Khokholov Yu. Building acoustic models for a large vocabulary continuous speech recognizer for Russian // In Proc. SPECOM' 2006, St. Peterburg, 2006, pp. 83-87.

[8] Кодзасов С.В., Кривнова О.Ф. Общая фонетика. – М.: РГГУ, 2001. – 592 с.

[9] Рассел С., Норвиг П. Искусственный интеллект: современный подход – М., 2006. – 1408 с.

[10] Карпов О.Н. Технология построения устройств распознавания речи. – Д., 2001. – 184 с.

[11] Kopeček I. Speech synthesis based on the composed syllable segments. // Proc. of the First Workshop on Text, Speech and Dialogue – TSD'98, 1998, pp. 259-262.

[12] 1. Карпов О.Н., Савенкова О.А. Некоторые эксперименты по повышению надежности распознавания слов заданного словаря // Системные технологии. Рег. межвуз. сб. науч. тр. Выпуск 6 (35), Днепропетровск, 2004, с. 60-66.

[13] Карпов О.Н., Савенкова О.А. Эвристический алгоритм поиска оптимальных решений сегментно-слогового синтеза // Искусственный интеллект. – 2007. - №4. – С. 378-385.

[14] Карпов О.Н. Вычислительные схемы представления функций многих переменных в классах функций меньшего числа переменных: Аналитическое описание поверхностей и спектров речевых сигналов. Моногр. – Д., 2003. – 120 с.