Сегментно-цілісна модель розпізнавання мовних образів на основі уявлень про функціональну асиметрію мозку при сприйнятті мовлення

Федяєв О.І., Бондаренко І.Ю.

Кафедра прикладної математики і інформатики
Донецький національний технічний університет, Україна
fedyaev@r5.dgtu.donetsk.ua

Анотація

Стаття присвячена дослідженню і реалізації двоканальній сегментно-цілісній моделі розпізнавання мовних образів. Теоретичною основою даної біонічної моделі є концепція про функціональну асиметрію півкуль головного мозку стосовно до процесу сприйняття усного мовлення. Пропонується структура системи розпізнавання мовних образів на основі взаємодії сегментної і цілісної підсистем розпізнавання, перша з яких реалізована в нейромережевому, а друга – у нечіткому базисі. Розглядаються методи узгодження рішень цих підсистем.

1. Вступ

Існує два підходи до рішення задачі автоматичного розпізнавання усного мовлення: інформаційний (прагматичний) і біонічний.

Перший підхід поєднує в собі багато методів, загальною рисою яких є феноменологічне моделювання процесу розпізнавання мови без урахування структури біологічного прототипу – слухової системи людини. Одними з найбільш популярних методів у рамках інформаційного підходу є КДП-метод, в основі якого лежать принципи динамічного програмування [1], та метод прихованих Марківських моделей [2]. Так, останній метод застосовується в багатьох сучасних системах розпізнавання усної мови. Однак припущення про імовірностну організацію мовних процесів у людини, що лежить в основі методу прихованих Марківських моделей, не є очевидним. Деякі практичні перевірки точності роботи систем розпізнавання, у яких застосовується даний метод, не дають результатів кращих, чим з п'ятьма відсотками помилок. Більш того, для текстів, що спонтанно вимовлені, імовірність правильного проголошення слів не перевищує однієї третини [3].

Другий підхід, біонічний, заснований на структурному моделюванні відділів центральної нервової системи, що здійснюють аналіз і сприйняття мовних сигналів. Але варто помітити, що сучасний рівень досягнень біологічних наук не забезпечує вичерпних знань в галузі фізіології мовлення і слуху. З погляду авторів, перспективним представляється рішення задачі розпізнавання мови в рамках об'єднання інформаційного і біонічного підходів, коли відтворення ряду відомих структур і механізмів діяльності головного мозку по сприйняттю мовлення поєднується з феноменологічним моделюванням недостатньо вивчених фізіологічною наукою елементів слухової системи людини.

У даній статті описується розробка системи розпізнавання мовних образів, заснованої на уявленнях фізіологів про функціональну асиметрію півкуль головного мозку. У процесі сприйняття ця асиметрія виявляється в тому, що мозок працює як двоканальна сегментно-цілісна система розпізнавання усного мовлення. Канали сегментного і цілісного розпізнавання, що відповідають лівій і правій півкулям, діють паралельно, забезпечуючи високу швидкість і надійність функціонування всієї системи в цілому [4]. У даній статті, що є подальшим розвитком роботи [5], пропонується реалізація першого каналу у вигляді нейромережевого пофонемного модуля розпізнавання, а другого – у вигляді нечіткого класифікатора цілісних образів (слів). Також розглянута задача формування колективного рішення про остаточне розпізнавання мовного образа на основі інтеграції результатів роботи каналів.

2. Структура двоканальної системи розпізнавання мовних команд

В основу двоканальної системи розпізнавання мовних образів покладені сучасні уявлення про механізми мовної діяльності людини [4]. Структурна схема роботи двоканальної системи розпізнавання представлена на рис.1.

У вхідному звуковому сигналі визначаються границі мовної ділянки – мовного образу, що передбачається – на основі функцій короткочасної енергії сигналу, числа переходів через нуль і кількості точок сталості. Далі виділений образ паралельно аналізується сегментним і цілісним каналами. Сегментний канал заснований на методі нейромережевої апроксимації фонем [6], а цілісний канал – на методі нечіткого зіставлення образів з оптимальним часовим вирівнюванням [7]. Ці канали формують незалежні набори слів-претендентів, тобто слів, до кожного з яких з визначеним коефіцієнтом упевненості може бути віднесений мовний образ, що розпізнається.

На останньому рівні системи, використовуючи набори слів-претендентів і відповідних їм коефіцієнтів упевненості, проводиться узгодження наближених рішень сегментного та цілісного каналів і приймається остаточне рішення про мовний образ, що розпізнається.

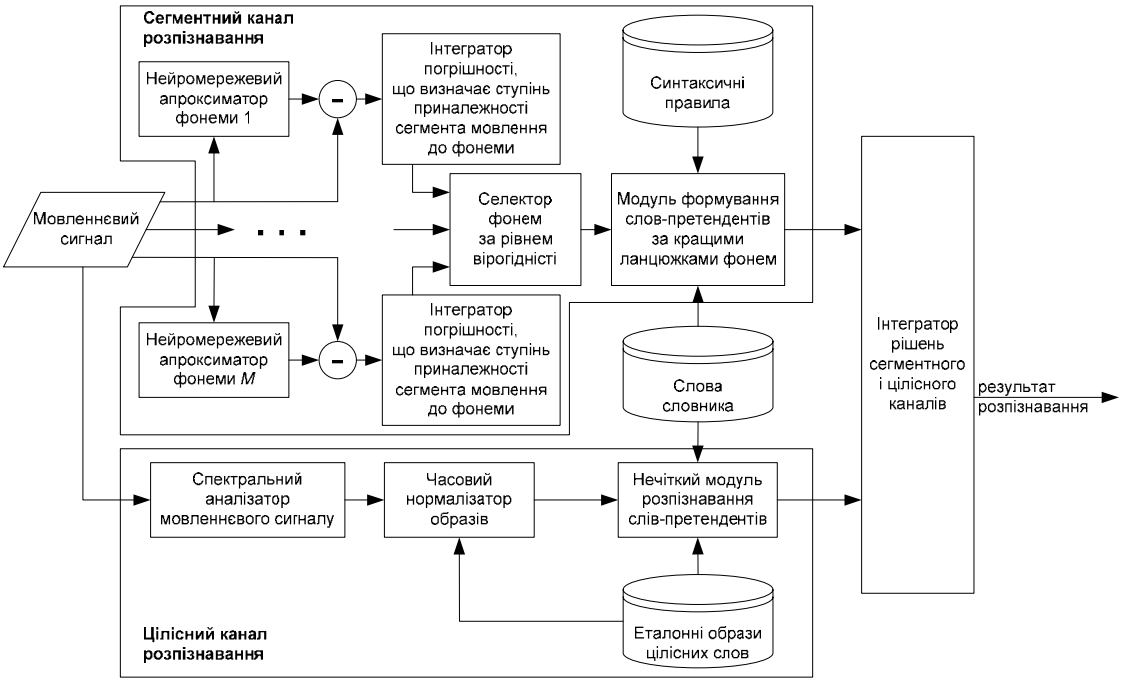

Рисунок 1: Структурна схема сегментно-цілісної системи розпізнавання мовних образів

3. Цілісний канал розпізнавання

Для розробки цілісного каналу розпізнавання запропонований метод нечіткого зіставлення образів з оптимальним часовим вирівнюванням, який є подальшим розвитком методу нечіткого зіставлення образів [8]. Застосування процедури оптимального часового вирівнювання для нормалізації мовних образів, що зіставляються, уздовж осі часу дозволяє, з одного боку, підвищити якість розпізнавання (у порівнянні з лінійним вирівнюванням), а з іншого, зменшити обсяг обчислень (у порівнянні з вирівнюванням на основі динамічного програмування) [7].

Як одиниці мовлення розглядаються слова, набір яких визначає словник цілісного каналу розпізнавання. Мовний сигнал представляється у вигляді двійкового двомірного спектрально-часового образу (ДСЧО), що дозволяє виділити місце розташування резонансних частот (локальних викидів), які є визначальною особливістю мовного сигналу [8].

ДСЧО мовного сигналу розглядається як бінарне відношення між множиною F (номерів частот f) і множиною T (номерів часових інтервалів t). Шляхом усереднення ДСЧО навчальних мовних сигналів заданого класу формується бінарне нечітке відношення, характерне для цього класу, у вигляді

$$f \in F, t \in T: \; F \; R \; T,$$

де R – нечітке відношення, що ставить у відповідність кожній парі елементів $(f,t) \in F \times T$ значення функції приналежності $\mu_R(x, y) \in [0,1]$. Набір нечітких відносин $R = \{r_1, r_2, \ldots, r_n\}$ визначає словник розпізнавання.

Для ДСЧО вхідного мовного сигналу y обчислюються ступені подібності S_j з кожним нечітким відношенням r_j, і як результат розпізнавання приймається такий номер j слова в словнику, що $j = \max_{j \in [1,n]} \{S_j\}$, де

$$S_j = \frac{\int r_j(f,t) \wedge y(f,t) df dt}{\int (\neg \, r_j(f,t)) \wedge y(f,t) df dt}.$$

Нормалізація образів уздовж осі часу перед їхнім зіставленням здійснюється за допомогою алгоритму оптимального часового вирівнювання. Задача оптимального часового вирівнювання розглядається як задача нелінійної оптимізації функції цілочисельного аргументу (числа часових інтервалів, що додаються до образа меншої довжини), а критерієм оптимальності вирівнювання є максимізація ступеня подібності образів, що зіставляються [7].

4. Сегментний канал розпізнавання

Сегментний підхід до розпізнавання мовлення заснований на фонетичному аналізі мовного сигналу. Використано метод нейромережевої апроксимації фонем, заснований на визначенні міри подібності фрагмента мовного сигналу до кожної з фонем і наступному виборі найбільш достовірного фонетичного ланцюжка [6]. Метод дозволяє з деякою погрішністю встановити, чи є фонема, що описується $F_k(t)$, фрагментом висловлення $A_w(t)$, де $A_w(t)$ – акустична форма висловлення w; $F_k(t)$ – акустична форма деякої фонеми. З цією метою функція $F_k(t)$ на відрізку $[t_0, t_1]$ представляється у виді множини пар

$$\{(X'(t), Y'(t))\}, \qquad (1)$$

де $X'(t)=(F_k(t-m),F_k(t-m+1),\ldots F_k(t-1))$, $m=const$; $Y'(t)=F_k(t)$; $t_0 \le t \le t_1$. Функція $A_w(t)$ представляється аналогічно у вигляді множини пар $\{X(t), Y(t)\}$.

Представлення $F_k(t)$ у вигляді (1) дозволяє сформувати нейромережеву функцію *NET*: $NET(X'(t))=Y'(t)$. Тоді міра відмінності Err_k ділянки $A_w(t)$ при $t \in [t_н, t_к]$ від $F_k(t)$ визначається за формулою

$$Err_k(t) = \left| Y(t) - NET(X(t)) \right|.$$

Таким чином, формується новий параметричний опис вихідного сигналу:

$$A_w(t) \rightarrow (Err_1(t), Err_2(t) \ldots Err_n(t)),$$

де $Err_k(t)$ – міра відмінності ділянки сигналу $A_w(t)$ від *k-ій* фонеми на фрагменті сигналу тривалості *m*.

Новий параметричний опис вихідного сигналу має переваги, пов'язані з більш високою стабільністю опису на стаціонарних ділянках, а також з можливістю однозначної інтерпретації отриманих величин. Однак складна форма і значна нестабільність мовного сигналу не дозволяють зробити висновок про фонему за окремими миттєвими значеннями мір відмінності $Err_k(t)$. Тому результати розпізнавання усереднювались на досить великій ділянці часу. Отриманий параметричний опис сигналу використовується при подальшій контекстній обробці, як це показано на схемі розпізнавання (рис. 1).

Перший рівень схеми складається з набору нейронних мереж, кожна з яких навчена розпізнаванню окремої фонеми. Виходи мереж інтерпретуються як прогноз наступних значень сигналу за умови, що має місце відповідна фонема. На другому рівні помилка прогнозу накопичується на всій довжині вікна сегмента мови. Інтегральна помилка надходить на третій рівень, де з усіх фонем вибираються найкращі. Отриманий набір бере участь у формуванні фонетичних ланцюжків, що являють собою гіпотези про слово, що вимовляється. Вимовлене слово визначається за ланцюжком з найбільшим ступенем вірогідності.

5. Інтегратор рішень сегментного і цілісного каналів

На останньому етапі роботи системи розпізнавання виникає необхідність у формуванні колективного рішення на основі інтеграції результатів роботи нечіткого і нейромережевого каналів. Для цього вводиться інтегратор, що формує підсумкове рішення в умовах ненадійної інформації.

Блок інтегратора реалізує метод неточних міркувань на основі фактора впевненості, який запропонований для системи MYCIN [9]. Вибір методу обумовлений його простотою і доброю спроможністю до компромісу при об'єднанні думок незалежних експертів (каналів розпізнавання). Блок інтегратора, представлений на рис.2, здійснює прямий логічний вивід на основі тверджень каналів розпізнавання з коефіцієнтами упевненості *CF* (Certainty Factor) і знань про формування колективного рішення групи незалежних експертів.

У задачах колективного розпізнавання з ненадійними даними важливу роль грає комбінований зв'язок, що позначається як КОМБ [9]. Він незалежно підкріплює або спростовує висунуту гіпотезу про вимовлене слово на підставі двох і більше думок експертів. Тому знання, представлені у виді продукційних правил (рис. 3), передбачають комбінований зв'язок між рішеннями, які запропоновані каналами розпізнавання. Припустимо, що один з експертів кожного каналу вже визначив ступінь надійності *X* і *Y* як результат попереднього розпізнавання, і необхідно зробити висновок (обчислити ступінь надійності висновку *V*), використовуючи правила з бази знань інтегратора рішень. Прямий логічний вивід підсумкового висновку заснований на використанні відомої методики оцінки передумови правил (антецедента) і врахування зв'язку КОМБ за методом MYCIN. Ступінь надійності поширюється по ієрархічній мережі логічного виводу (рис. 4), яка утворена продукційними правилами.

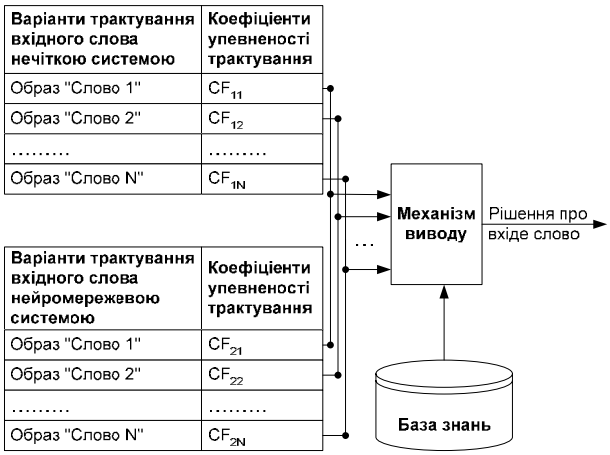

Рисунок 2: Експертний висновок по рішеннях каналів розпізнавання

1. ЯКЩО Трактування нейромережевою системою вхідного образа = образ «Слово 1»
   ТО Версія 1 = «Слово 1»
2. ЯКЩО Трактування нейромережевою системою вхідного образа = образ «Слово 2»
   ТО Версія 2 = «Слово 2»

. . . . . .

8. ЯКЩО Трактуваня нечіткою системою вхідного образа = образ «Слово 1»
   ТО Версія 1 = «Слово 1»
9. ЯКЩО Трактування нечіткою системою вхідного образа = образ «Слово 2»
   ТО Версія 2 = «Слово 2»

. . . . . .

Рисунок 3: База знань

У методі MYCIN ненадійність фактів представляється коефіцієнтом упевненості *CF*, що приймає значення від +1 (якщо факт свідомо правдивий) до 0 (для свідомо помилкових фактів). Запис *CF*[*V*, *Z*] будемо трактувати як коефіцієнт впевненості в істинності висновку *V*, якщо задовольняється передумова *Z* (консеквент). У процесі виводу при наявності зв'язку КОМБ окремо обчислюються $CF[V_i, X_i]$ і $CF[V_i, Y_i]$ за формулою:

$$CF[V, Z] = CF_{\text{правила}} \cdot CF_{\text{предпосылки}},$$

де $Z \in \{X, Y\}$; $CF_{\text{правила}} = 1$; $CF_{\text{предпосылки}} = CF_{ij}$.

Об'єднання рішень експертів по відповідних словах словника здійснюється за допомогою комбінованої функції

$$CF[V_i(X_i, Y_i)] = CF[V_i, X_i] + CF[V_i, Y_i] - CF[V_i, X_i] \cdot CF[V_i, Y_i].$$

Результат розпізнавання (номер розпізнаного слова) формується як номер максимального елемента в масиві $CF[V_i]$ (рис.4).

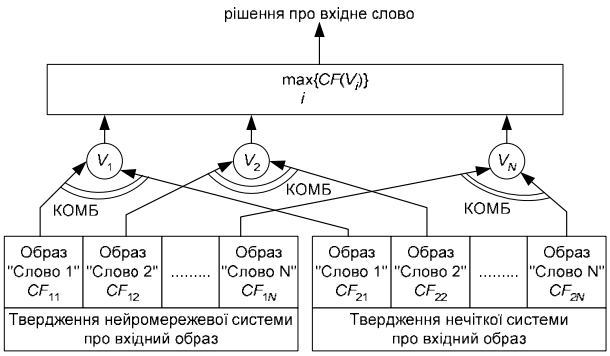

Рисунок 4: Схема виводу з урахуванням ненадійності знань

На рис.5 наведені результати колективного розпізнавання слова “Маркери” у зіставленні з іншими словами словника. У даному прикладі сегментний (нейромережевий) канал явно не віддає переваги якомусь одному слову, а з урахуванням думки іншого експерта – цілісного (нечіткого) каналу – інтегратор виробляє колективне рішення, що більш впевнено ідентифікує вхідне слово як “Маркери”.

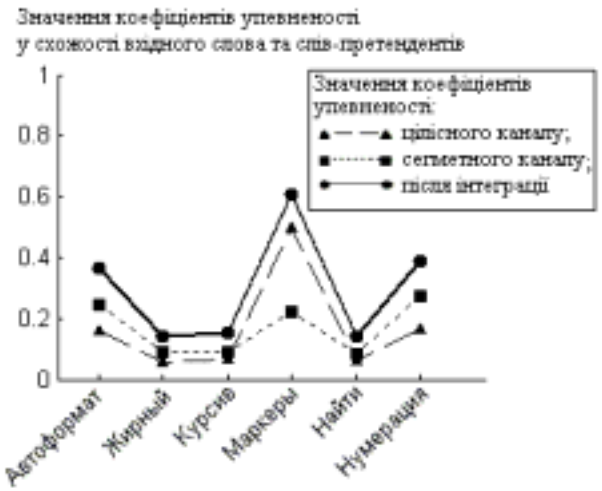

Рисунок 5: Результати колективного розпізнавання слова «Маркери» (словник на російській мові)

6. Висновки

Розроблено загальну структуру сегментно-цілісної системи розпізнавання мовних образів, заснована на фізіологічних уявленнях про функціональну асиметрію мозку при сприйнятті мовлення.

Розглянуто можливі реалізації сегментного і цілісного каналів. Гарні результати були отримані в результаті застосування методу нейромережевої апроксимації фонем для моделювання сегментного каналу розпізнавання, а методу нечіткого зіставлення образів з оптимальним часовим вирівнюванням – для моделювання цілісного каналу розпізнавання.

Запропоновано спосіб узгодження рішень сегментного і цілісного каналів у двоканальній моделі, який заснований на методі коефіцієнтів упевненості. Застосування цього методу дозволило врахувати колективну думку каналів розпізнавання як незалежних експертів і прийняти більш точне рішення про слово, що розпізнається. Об'єднання рішень каналів здійснюється інтегратором, що уявляє собою експертну систему, яка виконує логічний вивід рішення на основі ненадійних знань. Експерименти показали, що даний спосіб дозволяє ефективно вирішувати конфліктні ситуації, які виникають при розбіжності думок експертів (каналів розпізнавання).

7. Література

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

[2] Рабинер Л.Р. Скрытые Марковские модели и их применение в избранных приложениях при распознавании речи: обзор // ТИИЭР, т.77, №2, февраль 1989, С. 86-120.

[3] Галунов В.И., Соловьёв А.Н. Современные проблемы в области распознавания речи // Информационные технологии и вычислительные системы, №2, 2004 г. – С.41 – 45.

[4] Восприятие речи: вопросы функциональной асимметрии мозга / Морозов В.П., Вартанян И.А., Галунов В.И. и др. – Л.: Наука, 1988. – 135 с.

[5] Бондаренко И.Ю., Гладунов С.А., Федяев О.И. Сегментно-целостная структура канала речевого управления программными системами // Сб. трудов X нац. конференции по искусств. интеллекту с междунар. участием КИИ-2006. – М.: Физматлит, 2006. – с. 841 – 849.

[6] Гладунов С.А. Аппаратно-программные средства раздельной локализации фонем в системах речевого взаимодействия человека с ЭВМ: Автореф. дис…канд.техн.наук: 05.13.13 / ДонНТУ. – Донецк, 2005. – 22 с.

[7] Федяев О.И., Бондаренко И.Ю. Нечёткое сопоставление образов с оптимальным временным выравниванием для однодикторного и многодикторного распознавания изолированных слов // Сб. науч. трудов Донецкого нац. техн. ун-та. Серия «Информатика, кибернетика и вычислит. техника». 2007. Выпуск 8 (120). С.273–281.

[8] Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи и др. Распознавание речи // Прикладные нечёткие системы: Пер. с япон. Под ред. Т.Тэрано, К. Асаи, М. Сугено. – М.: Мир, 1993. – с. 157-170.

[9] Представление и использование знаний: Пер. с япон./ Под ред. Х.Уэно, М.Исидзука. – М.: Мир, 1989. – 220с.