УДК 681.3.06

[1]Л.О. Терейковська, [2]І.А. Терейковський

[1]Київський національний університет будівництва і архітектури, Київ
[2]Національний технічний університет України "Київський політехнічний інститут",Київ

ПРОБЛЕМА ГОЛОСОВОЇ ВЗАЄМОДІЇ В ДИСТАНЦІЙНОМУ НАВЧАННІ ВИЩОГО НАВЧАЛЬНОГО ЗАКЛАДУ

Проведено аналіз підходів вирішення проблеми голосової взаємодії в інформаційній системі дистанційного навчання. Сформульовано перелік актуальних задач, пов'язаних з означеною проблемою. Окреслені перспективні шляхи їх вирішення.

Ключові слова: дистанційне навчання, голосова взаємодія, спектральний аналіз, нейронна мережа

Проведен анализ подходов решения проблемы голосового взаимодействия в информационной системе дистанционного обучения. Сформулирован перечень актуальных задач, связанных с указанной проблемой. Очерчены перспективные пути ее решения.

Ключевые слова: дистанционное обучение, голосовое взаимодействие, спектральный анализ, нейронная сеть

The analysis of the approaches address the voice interaction in the information system of distance education. Formulated a list of tasks relevant tasks related to this problem . Outlines promising ways of solving it.

Keywords: distance learning, voice interaction, spectral analysis, neural network

Постановка проблеми

У загальному випадку дистанційна освіта являє собою комплекс освітніх послуг, що надається слухачам за допомогою спеціалізованого інформаційно-освітнього середовища, яке базується на засобах дистанційного обміну навчальною інформацією. Стосовно вищого навчального закладу, інформаційна система дистанційної освіти повинна забезпечити ефективну взаємодію територіально віддалених викладачів, навчальних матеріалів та слухачів (студентів). У більшості таких систем в якості основного комунікаційного засобу використовуються комп'ютерні мережі. Використання мережевих засобів комунікації, крім іншого, дозволяє підвищити якість навчання за рахунок занять, проведення яких доцільно реалізувати в реальному масштабі часу – дискусій, тестуванні, консультацій. Слід зазначити, що можливості загальнодоступних мережевих засобів дають змогу підвищити ефективність проведення таких занять за рахунок голосової взаємодії між трьома основними компонентами системи дистанційної освіти – викладачем, навчальними матеріалами та студентом. При цьому в розповсюджених системах дистанційної освіти ("Moodle", "Прометей", "SharePointLMS", "Микротест", "ИнтраЗнание" та "Батисфера") інструментальні засоби голосової взаємодії не відповідають висунутим вимогам. Означені передумови визначають основну проблему даної статті – підвищення ефективності інформаційної технології голосової взаємодії системи дистанційного навчання вищого навчального закладу .

Аналіз останніх досліджень і публікацій

Як правило, голосову взаємодію в процесі дистанційного навчання слід застосовувати під час проведення лекцій, семінарів, консультацій, лабораторних та практичних занять. Цілком природним доповненням до наведеного переліку є застосування голосової взаємодії для аутентифікації користувачів. Зазначимо, що під голосовою взаємодією будемо розуміти взаємодію між компонентами системи дистанційного навчання, яка базується на розпізнаванні голосових сигналів. У багатьох випадках таке розпізнавання має проводитись в автоматичному режимі. Наприклад, під час комп'ютерного тестування, коли студент повинен надати голосову відповідь на поставлене запитання. При цьому система розпізнавання має одночасно вирішити дві задачі:

1. Визначити правильність відповіді.
2. Провести аутентифікації користувача, тобто

визначити, чи насправді із системою взаємодіє заявлений студент.

Вирішення обох завдань базується на розпізнаванні голосової інформації користувача. Саме складність розпізнавання і визначає появу проблеми голосової взаємодії в інформаційній технології дистанційного навчання вищого навчального закладу. Зазначимо, що в загальному випадку таке розпізнавання полягає у послідовному вирішенні двох завдань:

1. Створенні формального опису голосової інформації.
2. Проведенні семантичного аналізу отриманого формального опису.

Досить часто під формальним описом голосової інформації розуміють його текстове представлення. При цьому сучасні теоретичні напрацювання в галузі семантичного аналізу текстової інформації не дозволяють створювати високонадійні інструментальні засоби, а в багатьох випадках відповідь та особу студента можна визначити на основі виявлення (відсутності) в ній декількох певних слів. Таким чином, стосовно інформаційної технології дистанційного навчання, розпізнавання голосової інформації зводиться до пошуку в ній ключових слів та розпізнавання диктора. Відповідно загальноприйнятий алгоритм пошуку ключових слів та розпізнавання диктора складається з таких етапів [1; 2]:

1. Оцифрування еталонного та піддослідного аудіосигналу.
2. Фільтрація шумів.
3. Виділення із сигналу окремих слів.
4. Обробка оцифрованого сигналу з метою зменшення обсягу вхідних даних системи розпізнавання.
5. Додаткова фільтрація спектру.
6. Стиснення спектру з метою врахування особливостей сприйняття звуку людиною та зменшення кількості вхідних параметрів системи розпізнавання.
7. Порівняння еталонного та піддослідного сигналів.

Разом з тим, методи реалізації вказаних етапів досить різноваріантні, базуються на різних математичних теоріях та адаптовані до різних умов, що визначає їх різну ефективність при використанні в інформаційній технології голосової взаємодії дистанційного навчання.

Формулювання мети статті

З точки зору підвищення ефективності інформаційної технології голосової взаємодії дистанційного навчання вищого навчального закладу визначити перспективні шляхи вирішення задачі пошуку ключових слів та розпізнавання диктора.

Виклад основного матеріалу дослідження

Для визначення перспективних шляхів пошуку ключових слів та розпізнавання диктора проведено аналіз окремих етапів наведеного алгоритму вирішення вказаних задач. Аналіз проводився з позицій застосування алгоритму в інформаційній технології дистанційного навчання.

На вхід сучасних систем розпізнавання аудіо-інформації, як правило, подається вже оцифрований сигнал. Оцифрування реалізується за допомогою стандартного обладнання персонального комп'ютера – звукової карти та мікрофону. При використанні загальнопоширеного обладнання дискретизації сигналу знаходиться в межах від 8000 до 44000 Гц. При цьому аналоговий вхідний сигнал квантується, тобто представляється у вигляді шістнадцяти розрядних або тридцяти двох розрядних чисел. Очевидно, що при застосуванні в системі дистанційного навчання слід орієнтуватись на деякі стандартні параметри оцифрування. Це можуть бути параметри, які відповідають найслабкішій конфігурації звукового обладнання: частота дискретизації 8000 Гц та шістнадцятирозрядне квантування.

Першочергова фільтрація шумів у дискретизованому сигналі полягає в накладенні на цей сигнал вікон різного типу – Кайзера, Хемінга та інших. В доступній літературі не наведено критеріїв, за якими вибирається тип вікна.

Після фільтрації шумів для виокремлення із звукового потоку окремих слів застосовується аналіз енергії сигналу протягом кожних 10-20 мс. Крім того, визначити початок/кінець слова можна за всплеском/затуханням величини сигналу.

Обробка вхідного оцифрованого сигналу з метою зменшення обсягу вхідних даних полягає у застосуванні різних методів спектрального аналізу даних. Спектральний аналіз даних реалізується або за допомогою віконного дискретного перетворення Фур'є або за допомогою дискретних вейвлет-перетворень. Дискретне перетворення Фур'є передбачає представлення ряду у вигляді декількох процесів, що складаються із синусоїд та косинусоїд різних частот [2]. Для цього оцифровані дані необхідно розвинути в ряд Фур'є:

$$\overline{X}(t) = a_0 + \sum_{i=1}^{q} (a_i c_i(t) + b_i s_i(t)) + e(t), \qquad (1)$$

де $c_i(t) = cos(2\pi f_i t)$, $s_i(t) = sin(2\pi f_i t)$, a_0 – коефіцієнт; a_i, b_i – коефіцієнти регресії, що вказують на ступінь кореляції функцій $c_i(t) = cos(2\pi f_i t)$ та $s_i(t) = sin(2\pi f_i t)$ з статистичними даними; $f_i = i/T$ – і-та гармоніка основної частоти 1/T; $q = (T-1)/2$; T – кількість точок ряду; $e(t)$ – випадкова складова.

Вираз $2\pi f_i$ позначають як λ_i та називають круговою частотою. Для розв'язання (1) необхідно підігнати $a_i c_{i,t}$ та $b_i s_{i,t}$ до статистичних даних

$$a_i = \frac{2}{T}\sum_{t=1}^{T}(\overline{X}(t)c_i(t)), \quad i = 1,2,...,q, \tag{2}$$

$$b_i = \frac{2}{T}\sum_{t=1}^{T}(\overline{X}(t)s_i(t)), \quad i = 1,2,...,q\,. \tag{3}$$

Після цього необхідно побудувати періодограму

$$I(f_i) = 0{,}5 \times T \times (a_i^2 + b_i^2), \quad i = 1,2,...,q\,. \tag{4}$$

Значення періодограми (I) на певній частоті (періоді) називають інтенсивністю й інтерпретують як дисперсію (варіацію) даних на цій частоті (періоді). Максимальна кількість гармонік (циклів), яку можна виділити із ряду в N точок, дорівнює $0{,}5N$, а мінімальний період, який можна виділити в процесі, не може бути менший за мінімальний період між двома послідовними реєстраціями статистичних даних. До переваг методу Фур'є належать апробованість та інтерпретованість результатів, а до недоліків – велика кількість обчислювальних операцій та неможливість визначення моментів виникнення частот, що призводить до помилок аналізу нестаціонарних голосових сигналів. Для виправлення недоліків використовують швидке віконне перетворення Фур'є, особливістю якого є застосування на інтервалах довжиною 10-20 мс, в межах яких сигнал залишається стаціонарним. Однак використання віконного перетворення значно ускладнює математичне забезпечення та може бути причиною неправильного розрахунку низькочастотних характеристик.

Основною перевагою спектрального аналізу методом вейвлет-перетворень є можливість не тільки визначити спектр сигналу, але й локалізувати частотні характеристики спектру у часі.

У загальному випадку неперервне вейвлет-перетворення функції $f(t)$ з кінцевою енергією у просторі $L^2(R)$ записується так:

$$W(a,b) = |a|^{-0,5}\int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt, \tag{5}$$

де W – коефіцієнт вейвлет-перетворення; ψ – базовий вейвлет (базисна функція); * – процедура комплексного спряження; a – масштаб вейвлета; b – зсув вейвлета, $a,b \in R$, $a \neq 0$.

При цьому функція ψ повинна відповідати таким вимогам: нульовий момент функції має дорівнювати нулю, енергія функції повинна бути кінцевою, концентруватись всередині деякого фінітного інтервалу та швидко зменшуватись поза вказаного інтервалу [3]. Для аналізу рядів з поліноміальним трендом у базисних вейвлетах центральні моменти v-го порядку мають дорівнювати нулю.

Особливістю дискретного вейвлет-перетворення неперервної функції є використання дискретних значень масштабу та зсуву вейвлету. Зазвичай, вказані величини задаються у вигляді степеневих функцій виду:

$$a = a_0^{-m}, \tag{6}$$

$$b = k \times a_0^{-m}, \tag{7}$$

де m – параметр масштабу; k – параметр зсуву; a_0 – початковий масштаб; m, k, a_0 – цілі числа, причому $a_0 > 1$.

З врахуванням (7), (8), вираз (5) запишемо так:

$$W(m,k) = |a_0|^{0,5m}\int_{-\infty}^{\infty} f(t)\psi^*(a_0^m \times t - k)dt. \tag{8}$$

Досить часто a_0 беруть таким, що дорівнює 2. Таке дискретне вейвлет-перетворення називають діадним Для діадного вейвлет-перетворення вирази (6), (7), (8) трансформуються так:

$$a = 2^{-m}, \tag{9}$$

$$b = k \times 2^{-m}, \tag{10}$$

$$W(m,k) = 2^{0,5m}\int_{-\infty}^{\infty} f(t)\psi^*\left(2^m \times t - k\right)dt. \tag{11}$$

Процедура дискретного вейвлет-перетворення починається з початкового масштабу $a = a_0^{-m}$, якому відповідає рівень мінімально допустимої часової дозвільної здатності сигналу. Процедура продовжується разом з дискретним збільшенням масштабу за рахунок дискретного зменшення параметра m. Таким чином, аналіз починається з високих частот і здійснюється в бік низьких частот. Перше значення масштабу відповідає найбільш стиснутому вейвлету. При збільшенні величини масштабу вейвлет розтягується. При цьому розтягування вейвлету в a разів по горизонталі приводить до його зменшення в a разів по вертикалі.

На початку аналізу вейвлет розміщується на початку сигналу (t=0), помножується із сигналом, інтегрується на інтервалі свого визначення і нормалізується на $|a_0|^{0,5m}$. Результат обчислення $W(a,b)$ розміщується в точці ($a = a_0^{-m}, b = 0$) масштабно-часового спектру перетворення [2; 4]. Зсув b може розглядатись як час з моменту t=0, при цьому координатна вісь b повторює часову вісь функції. Після цього вейвлет масштабу a_0^{-m} посувається вправо на значення $b = k \times a_0^{-m}$ і процедура повторюється. На частотно-часовій

площині отримуємо значення, яке відповідає $t = b$ і $a = a_0^{-m}$. Процедура повторюється доти, поки вейвлет не досягне кінця сигналу. Для обчислення наступного масштабного рядка значення *a*, дискретно збільшується на деяке значення, визначене параметром *m*. Тим самим реалізується дискретизація масштабно-часової площини. Максимальне значення масштабу *a* відповідає тривалості всього аналізованого ряду даних. Для деталізації самих високих частот сигналу мінімальний розмір вікна вейвлету не має перевищувати періоду самої високочастотної гармоніки.

Якщо в сигналі наявні спектральні компоненти, які відповідають поточному значенню *a*, то інтеграл добутку вейвлета із сигналом на інтервалі, де ця спектральна компонента наявна, дає відносно велике значення. У протилежному випадку – добуток невеликий або дорівнює нулю, через те, що середнє значення вейвлет-функції дорівнює нулю.

Додаткова фільтрація спектру проводиться з метою виокремлення із нього частот, нехарактерних для людського голосу. Зазначимо, що діапазон частот, які чує людина знаходиться в межах від 16 до 20 000 Гц. Однак, в більшості систем, виходячи в тому числі й із можливостей комп'ютерної звукової апаратури, застосовується діапазон частот від 50 до 16 000 Гц.

На сьогодні загальноприйнято проводити стиснення спектру за допомогою процедури визначення мел-спектральних коефіцієнтів. Її результатом є 16-24 коефіцієнти, які достатньо повно характеризують весь діапазон звукових частот, які відчуває людина. Ще один підхід стиснення спектру базується на методі визначення формант голосового сигналу [1; 2].

Для порівняння еталонного та піддослідного сигналів натепер переважно використовуються сховані марковські моделі, методи динамічного програмування та нейронні мережі.

Використання схованих марковських моделей базується на постулаті, що голосовий сигнал може бути розділений на стаціонарні фрагменти, які відповідають окремим станам ланцюга Маркова

$$O = \{o_1, o_2, ..., o_\tau\}. \quad (12)$$

Перехід між станами відбувається миттєво, а ймовірність відображення породженого моделлю фрагменту залежить тільки від поточного стану моделі та не залежить під попередніх станів. Власне схованою марковською моделлю називається модель, що складається із *N* станів, в кожному із яких деяка система може приймати одне з *M* значень деякого параметра. Як правило, модель задається виразом

$$\lambda = \{A, B, \pi\}, \quad (13)$$

де *A* – матриця ймовірностей переходів по станах; *B* – вектор ймовірностей випадіння кожного із *M* значень параметра в кожному із *N* станів; π – вектор розподілу початкових ймовірностей.

Перший етап використання моделі полягає у її навчанні на прикладах (12), що відповідають еталону ключового слова. Результатом навчання є розрахунок параметрів виразу (13). Після цього на вхід моделі можна подавати послідовність (12), яка відповідає невідомому голосовому сигналу. Розв'язання задачі знаходження ймовірності появи цієї послідовності у кожній із попередньо навчених моделей дозволить визначити ту модель, яка найбільш достовірно відповідає голосовому сигналу, а значить, і розпізнати ключове слово. До основних недоліків схованих марковських моделей належить велика обчислювальна складність та складність формування бази даних ключових слів.

Застосування методів динамічного програмування зводиться до розрахунку ключового слова найбільш схожого на невідоме. Критеріями схожості слів виступають відстань Евкліда, відстань Хемінга та інші. В доступній літературі методики вибору критерію схожості не знайдено.

Використання нейронних мереж базується на їх здатності класифікувати голосові сигнали, задані за допомогою коефіцієнтів, які відповідають спектральним характеристикам [4]. Не зважаючи на перспективність даного напряму, застосуванню нейронних мереж перешкоджає відсутність методики оптимізації типу та параметрів мережі. У результаті можна зазначити, що основною невирішеною задачею в області пошуку ключових слів та розпізнавання диктора є задача порівняння еталонних та невідомих голосових фрагментів.

Висновки

1. Доведено, що проблема голосової взаємодії в дистанційному навчанні може бути зведена до розв'язання задач розпізнавання ключових слів та розпізнавання диктора.
2. Визначено недоліки та переваги основних методів обробки звукових сигналів в задачах розпізнавання ключових слів та розпізнаванні диктора.
3. Визначено перелік задач, розв'язання яких дозволить підвищити ефективність голосової взаємодії системи дистанційного навчання.

Список літератури

1. Гребнов С.В. Аналитический обзор методов распознавания речи в системах голосового управления / С.В.Грубнов // Вестник ИГЭУ. – 2009. – №3. – С.22-25.

2. Рабинер Л.Р. Цифровая обработка рече вік сигналов / Л.Р. Рабинер, Р.В. Шафер. – М.: Радио и свіязь. – 1981. – 496 с.

3. Терейковская Л.А. Разработка статистической модели расчета периодических составляющих динамики функциональных параметров Internet-серверов / Л.А Терейковская // Управління розвитком складних систем: зб. наук. праць. – 2010. – Випуск 3. – С. 107–111.

4. Терейковський І.А. Нейронні мережі в засобах захисту комп'ютерної інформації / І.А. Терейковський. – К. : ПоліграфКонсалтинг. – 2007. – 209 с.

Стаття надійшла до редколегії 21.03.2013

Рецензент: д-р техн. наук, проф. Ю. М. Тесля, Київський національний університет будівництва і архітектури, Київ.