Дослідження ефективності методу текстонезалежної ідентифікації диктора, що враховує фонетичні особливості звуків мовлення

Т.В. Єрмоленко, М.С. Клименко

Відділ розпізнавання мовних образів
Інститут проблем штучного інтелекту, Україна
Naturewild71@gmail.com, nk@xaker.ru

Анотація

У статті запропонована модифікація методу текстонезалежної ідентифікації диктора на основі гаусових сумішей. Запропонований метод формує комплексну модель диктора, елементи якої отримані в результаті обробки ділянок мовленнєвого сигналу, що належать різним фонетичним класам. Проведені дослідження показали, що розбиття акустичного простору голосу диктора на множину класів, які представляють деякі фонетичні події, приводить до збільшення ефективності ідентифікації по голосу.

1. Вступ

У даний час ведеться активне впровадження голосової біометрії в багатокористувацькі автоматизовані системи різного спектру застосування. Основна перевага голосової ідентифікації перед іншими біометричними системами полягає у можливості отримання і передачі біометричних даних у центр засвідчення без застосування спеціалізованих і дорогих знімачів біометричної інформації. Крім того, процес аутентифікації не вимагає від користувача безпосереднього контакту з елементами пропускної системи, що відкриває можливість проведення даної процедури віддалено, наприклад, через Інтернет або мережу мобільного зв'язку. Біометрична голосова аутентифікація є ефективним засобом і широко використовується для забезпечення безпеки даних.

Автоматичні системи розпізнавання диктора діляться на системи ідентифікації та верифікації. Ідентифікація являє собою процес порівняння мовленнєвого фрагменту із образами множини дикторів, зареєстрованих у системі. Користувач, не занесений в систему, буде ідентифікований як диктор, чий образ у більшій мірі відповідає вхідному вислову. У разі необхідності виявлення незнайомого диктора в систему додається універсальна фонова модель [1], створена із сукупності характеристик всіх зареєстрованих користувачів. Верифікація диктора - це процес прийняття або відхилення факту приналежності мовленнєвого фрагменту заявленому користувачеві. У даному випадку буде прийнятий лише той мовний сигнал, у якого значення відповідності з образом диктора буде не нижче порогового.

Таким чином, основною відмінністю між ідентифікацією і верифікацією є кількість альтернативних рішень, яка при ідентифікації дорівнює кількості множини образів дикторів, а при верифікації - тільки прийняттю або відхиленню, незважаючи на кількість зареєстрованих користувачів. Виходячи з цього слід зауважити, що ефективність ідентифікації диктора часто зменшується при збільшенні множини моделей, тоді як ефективність верифікації диктора близька до постійної величини. Отже, ідентифікація за мовленням висуває підвищені вимоги до роздільності моделей, у той час як для верифікації необхідна точність передачі індивідуальних характеристик кожного диктора окремо. Ускладнює задачу ряд факторів: нестаціонарність вимовляння, ефект реверберації голосу, а також спотворення і перешкоди в каналах зв'язку.

Крім того, спостерігається підвищений попит на створення текстонезалежних систем, спричинений як простотою використання, так і необхідністю застосування в правозахисній сфері. Але ефективність та швидкість розпізнавання таких систем на даний момент значно поступається текстозалежним аналогам, не забезпечуючи достатньо високої надійності розпізнавання дикторів. Тому актуальним завданням є створення алгоритмів, що підвищують точність текстонезалежної ідентифікації та зберігають при цьому прийнятні показники обчислювальної трудомісткості.

У даній роботі наведено чисельне дослідження модифікованого методу побудови моделі диктора, що описано в [2]. Модифікація полягає в поділі акустичного простору голосу диктора на широкі фонетичні класи (ШФК) шляхом попередньої дикторонезалежної сегментації мовленнєвого сигналу з одночасною класифікацією його сегментів. Модель диктора описується набором моделей, отриманих в результаті обробки акустичних характеристик на ділянках сигналу, що відповідають різним ШФК. Результати дослідження дозволять оцінити можливість практичного використання запропонованого підходу до побудови моделі диктора.

2. Методи побудови ознакових описів і прийняття рішень в задачах розпізнавання диктора

Обидві задачі розпізнавання по голосу спираються на моделі диктора - набори структурованих акустичних ознак, обчислених по мовленнєвому сигналу користувача. Структура таких моделей досить різноманітна і безпосередньо залежить від використовуваних акустичних характеристик і класифікаторів.

Індивідуальність мови диктора формується особливостями будови його мовного тракту і станом нервової системи, яка надає безпосередній вплив на процес артикуляційної діяльності. Акустичні характеристики зобов'язані передавати дані особливості

диктора, а також поєднувати в собі стійкість до спотворень різного роду і компактність представлення для можливості швидкої обробки, зберігання та порівняння еталонних значень.

Методи розв'язання обох задач використовують спектральні акустичні ознаки мови на основі Фур'є та вейвлет-спектра, кепстральних коефіцієнтів, а також їх похідних за часом у вигляді векторів дійсних чисел. До найчастіше використовуваних акустичних ознак можна віднести:

- мел-частотні кепстральні коефіцієнти (Mel Frequency Cepstral Coefficient – MFCC);
- перцептуальні коефіцієнти і кепстральні коефіцієнти лінійного передбачення.

Для врахування динамічної складової вектори моментальних характеристик, що обчислені на наборі послідовних вікон, можуть бути представлені у вигляді матриці [3].

З методів класифікації моделей найпоширенішими на даний момент є: векторне квантування, гаусові суміші та метод опорних векторів.

Метод ***векторного квантування*** має на меті поділ всього простору ознак на області, в яких сконцентровані акустичні ознаки диктора. Даний метод будує модель у вигляді набору векторів ознак, які є центроїдами кластерів, що не перетинаються між собою. Структура моделі може бути доповнена ваговими коефіцієнтами для посилення важливості окремих кластерів. Також можлива структура з веденням єдиної загальної карти кластерів, за якої моделі дикторів описуються кодовими книгами - статистичними даними про входження векторів ознак в кластери фонової моделі.

Моделі, створювані на основі ***гаусових сумішей***, продовжують ідею векторного квантування, але з тією різницею, що класи в просторі ознак описуються у вигляді багатовимірного імовірнісного розподілу. В якості функції розподілу зручно застосувати функцію Гауса. Оскільки емпіричний розподіл даних далекий від нормального, то необхідна точність досягається за рахунок подання його у вигляді зваженої суми M нормальних розподілів:

$$p\left(\bar{x} \mid \lambda\right) = \sum_{i=1}^{M} w_i \, p_i\left(\bar{x}\right),$$

де $\bar{x}$ - N-вимірний вектор ознак;

w_i - вага компонентів моделі;

p_i - багатовимірні функції щільності розподілу складових моделі.

Таким чином, повністю модель описується векторами математичного очікування, коваріаційний матрицями та вагами сумішей для кожного компонента моделі.

Широко використовуваним методом оцінки параметрів моделі є метод максимізації правдоподібності. Функція максимальної правдоподібності представлена у вигляді:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\bar{x}_t \mid \lambda), \qquad (1)$$

де T - кількість векторів ознак;

$X = \left\{\bar{x}_1, ..., \bar{x}_T\right\}$ - послідовність векторів ознак.

Оскільки (1) - нелінійна функція від параметрів моделі, її безпосереднє обчислення неможливо, тому оцінки параметрів можуть бути отримані ітераційно.

Виходячи з припущення, що всі диктори однаково ймовірні, спрощене правило класифікації має вигляд:

$$res = \arg \max_{1 \le k \le S} p\left(X \mid \lambda_k\right), \qquad (2)$$

де S - кількість дикторів.

Методом ***опорних векторів*** досягається знаходження в багатовимірному просторі ознак гіперплощини, що є рівновіддаленою від крайніх (опорних) векторів протилежних класів. Таким чином можливо виконати поділ тільки двох акустичних класів, а для більшої множини дикторів може бути використана схема «один проти кожного». У цьому випадку модель диктора складається з множини гіперплощин, кожна з яких відокремлює ознаки даного диктора від одного з решти. Це означає, що для системи, що зберігає N моделей дикторів, необхідна побудова матриці попарно розділяючих гіперплощин розмірністю NxN. Альтернативою є зворотнє рішення даної задачі, коли близькі моделі дикторів об'єднуються в групи. Процедура групування повторюється ітераційно і результатом є деревовидна структура, у вузлах якої знаходяться рівняння, що розділяють найближчі класи групових ознак, а листя розділяють безпосередньо акустичні ознаки дикторів. За великої кількості дикторів дане представлення значно зменшує обсяг збережених даних і кількість обчислень при порівнянні. Ефективність буде залежати від збалансованості бінарного дерева. Останнім часом найчастіше застосовується схема «один проти всіх», метою якої є виокремлення ознак конкретного диктора від всіх інших. Рішення у такому вигляді ідеально пристосоване для задачі верифікації, але так само широко застосовується і при ідентифікації диктора [4]. У випадку лінійної нероздільності для побудови гіперплощини між класами, що частково перетинаються, обмеження доповнюються скалярним параметром допуску. Іншим способом, що дозволяє розпізнавати лінійно-нероздільні класи, є відображення початкового простору ознак у простір більшої розмірності, в якому класи можуть бути розділені лінійно. Дане перетворення виконується за допомогою функції ядра. Параметри методу (скалярний параметр допуску та параметри ядра), як правило, визначають за допомогою перебору деякої множини значень.

3. Дослідження ефективності модифікованого методу ідентифікації по голосу, що враховує широкі фонетичні класи звуків

При поділі акустичного простору голосу диктора на ШФК стає можливим формування моделі диктора за акустичними ознаками звуків, близьких за способом формування, що дозволить створити більш точну модель.

У роботі використовувалися 4 класи звуків російської мови:

- *Voc* - голосні {[і], [е], [о], [у], [а], [и]};
- *Sh* - глухі приголосні {[ф], [с], [х], [ш], [ф’], [с’], [х’], [щ], [ц], [ч]};
- *Cons* - дзвінкі приголосні {[в], [з], [ж], [в'], [з'], [ж'], [б], [д], [г], [б'], [д '], [г']};
- *Son* - сонорні {[й], [л], [л '], [м], [н], [м'], [н '], [р], [р']}.

Крім звуків мови в якості п'ятого класу був використаний шум - фрагмент сигналу, що не містить мовлення.

Для ідентифікації застосовувався метод гаусових сумішей, в якості акустичних характеристик використані MFCC, що формують 13-вимірний вектор ознак (ВО), цієї кількості коефіцієнтів цілком достатньо для обробки мовних сигналів.

Мовленнєвий сигнал розбивався на фрейми довжиною близько 20мс з половинним перекриттям, далі проводилася процедура класифікації фреймів по ШФК. Моделі ШФК, за якими здійснювалася класифікація, будувалися на основі гаусових сумішей розмірністю 10, для навчання були використані записи дикторів (чоловіків і жінок з різними голосовими даними) загальною тривалістю близько 20 хвилин.

За набором ВО, отриманих на множині фреймів, що належать одному ШФК, виконувалася кластеризація методом *K*-середніх з ітеративним додаванням центроїдів (поділом кластера з максимальним радіусом на два). Кількість центроїдів, а отже, і розмірність гаусової суміші, визначається критерієм ефективності опису вибірки ICL-BIC без використання штрафу на число компонент. Остаточне позиціювання центроїдів виконується методом максимізації правдоподібності. Створення моделі диктора, що відповідає певному ШФК, завершувалося побудовою гаусової суміші з використанням отриманих центроїдів. Результуюча модель являє собою набір з 4 моделей диктора, сформованих для різних ШФК:

$$\lambda_k = \left(\lambda_k^{Voc}, \lambda_k^{Sh}, \lambda_k^{Cons}, \lambda_k^{Son}\right)$$

У чисельних дослідження брали участь 100 дикторів з різними голосовими даними. Для побудови моделей були записані фрагменти мовлення дикторів тривалістю 1 хвилина. Запис здійснювався динамічним мікрофоном у приміщенні без сторонніх шумів (рівень шуму -45dB) з частотою дискретизації 44,1 кГц і глибиною квантування 16 біт.

Для проведення порівняльного аналізу ефективності ідентифікації згідно методом, викладеним в [2] були отримані моделі дикторів, які не враховують розбиття по ШФК, а також моделі, навчені тільки на фреймах, що належать одному ШФК.

Однією з проблем при навчанні сумішей гаусових моделей є вибір числа компонентів моделі. У даній роботі авторами використовувався критерій ефективності опису вибірки сумішшю з k компонент, що включає в себе штраф на кількість компонент (критерій ICL-BIC), який описаний в [5]. Максимальні та мінімальні розмірності сформованих моделей дикторів, що було отримано у результаті чисельних досліджень за допомогою критерію ефективності ICL-BIC, представлені в таблиці 1.

При створенні моделей граничною була обрана розмірність 20. Як видно з результатів моделювання без урахування ШФК, дане значення було досягнуто при обробці кожного диктора. Цей факт свідчить про велику ентропію кластеризованих даних. При побудові моделей для кожного ШФК ситуація надмірної кластеризації спостерігалася тільки для класу *Voc*.

Таблиця 1. Максимальні та мінімальні кількості гаусових компонент у суміші, отриманих при побудові моделей дикторів за критерієм ефективності ICL-BIC

Тип моделі	MIN	MAX
Без урахування ШФК	20	20
Voc	11	20
Sh	8	19
Cons	6	11
Son	16	20

Слід зауважити, що було виконано побудову моделей з граничною розмірністю більшого порядку. Результатом стало як збільшення середньої розмірності до 27, так і прояв властивостей надмірної кластеризації внаслідок надлишкового поділу областей з великим внутрікластерним розкидом. Тому було прийнято рішення не підвищувати розмірність.

На відміну від правила класифікації (2), яке використовує функцію максимальної правдоподібності (1) в даній роботі з метою згладжування ефекту, виробленого викидами, що мають нульову або близьку до нуля оцінку ймовірності, замість добутку використовувалася сума (3), усереднена по T - кількістю фреймів мовленнєвого сигналу, за яким проводиться ідентифікація:

$$F(k) = \frac{\sum_{t=1}^{T} p(\bar{x}_t \mid \lambda_k)}{T} \qquad (3)$$

У цьому випадку правило класифікації приймає вигляд:

$$S = \arg \max_{1 \le k \le 100} F(k)$$

У даній роботі належність послідовності векторів ознак моделі диктора визначається аналогом функції правдоподібності (3). Її значення повинно бути максимальним, якщо модель і мовленнєвий сигнал, за яким отримана послідовність ВО, належать одному диктору. Дана залежність демонструє точність передачі характеристик диктора, а низький розкид одержуваних значень може свідчити про стабільність результатів. Було проведено порівняння послідовності векторів ознак мовленнєвих фрагментів кожного диктора із відповідною йому моделлю. Акустичні ознаки отримані із сигналів тривалістю не менше 5 сек, а за значеннями функції (3)

обчислені математичне сподівання (МС) та середньоквадратичне відхилення (СКВ).

При розгляді отриманих значень функції (3), відповідних моделям без урахування ШФК спостерігався найвищий розкид результатів. У порівнянні з цими даними МС для комплексних моделей зросло в середньому по всіх дикторам в 5 разів, в той час як СКВ знизилося на 15%.

Якщо провести порівняння моделей, що навчені на фреймах, які належать одному ШФК, з моделлю без урахування ШФК, то можна сказати наступне (табл. 2):

- значення МС і СКВ функції (3) для моделей, навчених на фреймах класу *Voc* порівняно з показниками моделей без урахування ШФК;
- для моделей, навчених на фреймах класів *Sh* і *Cons* МС значень функції (3) зросло в середньому по всіх дикторам в 2 рази, проте їх СКВ порівняно з СКВ для моделей без урахування ШФК;
- досліджувані показники для моделей, навчених на фреймах класу *Son*, є найкращими.

Проведемо аналогічне дослідження поведінки функції (3) для випадку, коли модель і сигнал, що підлягає ідентифікації, належать різним дикторам. В даному випадку значення функції (3) повинно бути якомога ближче до 0.

Отримані значення в 10-50 разів менше, ніж значення функції (3), обчислені для випадку, коли модель і сигнал, по якому проводиться ідентифікація, належать одному диктору. Відповідні дані наведені в таблиці 2.

Таблиця 2. Відношення статистичних параметрів значень функції (3) різних типів моделей до значень моделі без поділу на ШФК

Тип моделі	**МС, %**	**СКВ, %**
модель і сигнал, що підлягає ідентифікації, належать одному диктору		
Комплексна модель	508,6	74,1
Voc	87,3	93,7
Sh	192,4	104
Cons	201,6	86,3
Son	1088	107,1
модель і сигнал, що підлягає ідентифікації, належать різним дикторам		
Комплексна модель	97,7	144,3
Voc	99,2	118,3
Sh	101	214,2
Cons	95,1	396,7
Son	98,4	483,2

При проведенні ідентифікації за допомогою різних моделей дикторів результати, що було отримано, показали перспективність застосування комплексної моделі (табл. 3).

Таблиця 3. Показники ефективності ідентифікації при використанні різних типів моделей диктора

Тип моделі	**Імовірність ідентифікації, %**
Без урахування ШФК	94,8
Комплексна модель	98,7
Voc	94,3
Sh	96,7
Cons	95,1
Son	97,2

4. Висновки

У даній роботі було проведене дослідження ефективності методу ідентифікації, що використовує комплексну модель диктора, яка враховує ШФК. Аналізуючи отримані результати, можна зробити наступні висновки.

1. Використання комплексної моделі диктора, елементи якої отримані в результаті обробки ділянок сигналу, що належать різним ШФК, дозволило підвищити ймовірність ідентифікації більш, ніж на 3%.

2. Елементи комплексної моделі диктора, навчені на фреймах, що належать тільки одному ШФК, володіють різними розділовими здібностями в залежності від складу фонетичного класу, а отже, роблять різний внесок у ідентифікаційні властивості результуючої моделі.

Ефективність ідентифікації диктора можливо підвищити за рахунок:

- поліпшення роздільних властивостей ВО шляхом додавання до нього робастних акустичних характеристик, які володіють ідентифікаційними властивостями;
- додаванням в цільову функцію вирішального правила вагових коефіцієнтів, що відображають розділові здатності моделей, навчених на одному ШФК.

Розвитком даної роботи є порівняльне дослідження різних моделей формування ознакового опису мовних сигналів і систем розпізнавання дикторів з метою визначення перспективних напрямків їх створення.

5. Література

[1] Wei-Qiang Zhang and Jia Liu, “Discriminative Universal Background Model Training for Speaker Recognition”, Speech and Language Technologies: 241-256, 2011.

[2] Садыхов Р. Х., Ракуш В. В. Модели гауссовых смесей для верификации диктора по произвольной речи, доклады БГУИР, №4: 95-103, 2003

[3] Wu Q, Zhang L. Q. and Shi G. C., “Robust feature extraction for speaker recognition based on constrained nonnegative tensor factorization”, Journal of computer science and technology 25(4): 745-754.

[4] Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers // Advances in Kernel Methods. — MIT Press, Cambridge, USA, 1998.

[5] Сорокин В.Н. и Цыплихин А.И., “Верификация диктора по спектрально-временным параметрам речевого сигнала”, Информационные процессы. Т. 10, № 2, С. 87-104.