Експериментальні дослідження з адаптації до голосу диктора на базі корпусу АКУЕМ

О.А. Юхименко

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680
yukhymenko@uasoiro.org.ua

Abstract

This article is devoted to the problems of adaptation to new announcer voice for speech recognition systems. The results of three adaptation experiments based on AKUEM speech corpus are described. Comparison with previous experiments are discussed.

1. Вступ

До роботи з корпусом АКУЕМ була проведена серія експериментальних досліджень з адаптації, застосовані різні підходи [1,2]. Слід зазначити, що вони були проведені в рамках пофонемного послівного розпізнавання. Всі диктори, записи котрих використовували в експериментах, наговорювали визначені певні окремі слова, при цьому вимовляли їх достатньо чітко й в нормальному темпі. Розпізнавання було послівним. Канал зв'язку (мікрофон, акустика, фон тощо) був для всіх один й той самий та достатньо якісний. Словник використовувався невеликий (трохи більший за 2 тисячі слів). Кількість дикторів також була невеликою. В даній роботі автор представляє дані експериментальних досліджень, котрі були отримані в дещо інших умовах. Це спонтанне мовлення, про що піде мова нижче.

2. Лінійні перетворення при адаптації акустичних моделей

При створенні системи розпізнавання сигналів мовлення необхідно провести процедуру навчання розпізнаванню. При пофонемному розпізнаванні кожна фонема має свою акустичну генеративну модель, котра являє собою певну кількість станів з певними переходами між ними [1]. При цьому кожний стан моделі має свої ймовірнісні параметри – середній вектор спостереження $\mu = [\mu_1, \mu_2, ..., \mu_n]^T$ та матрицю коваріацій Σ розмірністю $n \times n$, де n – розмірність вектора первинних ознак сигналу. Ці μ та Σ є параметрами n-вимірного нормального закону розподілу. Стан моделі може задаватися декількома параметрами (парами), то тоді говорять, що стан описується сумішшю гаусіанів (нормальних розподілів). Проведення процедури навчання передбачає конкретне обчислення за допомогою ітераційних процедур саме цих ймовірнісних параметрів для всіх фонем в системі розпізнавання. Для двох систем розпізнавання, навчених на двох різних дикторів, ці ймовірнісні параметри будуть різнитися між собою, чим й пояснюється незадовільна точність розпізнавання якогось диктора на чужій системі.

Але цілком можливо обчислити лінійні перетворення, які переводять початкові середні вектори та матриці коваріацій опорного диктора або кооперативу дикторів у середні вектори та матриці коваріацій нового диктора. Ефектом цих перетворень є зсув середніх векторів моделей фонем та зміна дисперсій у початковій системі таким чином, що кожний стан у системі акустичних моделей фонем буде точніше генерувати дані адаптації.

Лінійне перетворення для середніх векторів записується у вигляді:

$$\hat{\mu} = W\xi, \quad (1)$$

де $\hat{\mu}$ – середній вектор нового диктора, W є матрицею розмірністю $n \times (n+1)$, ξ – середній розширений вектор опорного диктора,

$$\xi = [1, \mu_1, \mu_2, ..., \mu_n]^T. \quad (2)$$

Лінійне перетворення коваріаційних матриць записується у вигляді:

$$\hat{\Sigma} = H\ \Sigma H^T, \quad (3)$$

де H – матриця перетворення матриці коваріацій Σ опорного диктора, розмірність – $n \times n$.

Щоб покращити гнучкість процесу адаптації, можна визначити відповідну множину базових класів, яка залежатиме від кількості доступних адаптаційних даних [3]. Якщо доступна мала кількість адаптаційних даних, то тоді буде генеруватися загальне адаптаційне перетворення . Загальне перетворення застосовується до кожної компоненти гаусіанів в множині моделей. Одначе, якщо адаптаційних даних стає більше, то можливо покращити адаптацію шляхом збільшення кількості перетворень. Тоді кожне перетворення стає більш конкретним й застосовується до певної групи гаусіанів. Наприклад, гаусіани можуть бути згруповані в широкі фонетичні класи: пауза, голосні, назальні, фрикативні тощо. В цьому випадку адаптаційні дані повинні використовуватися для побудови більш конкретних перетворень широких класів, щоб застосувати ці перетворення до цих угруповань.

Зв'язування кожного перетворення через множину компонентів суміші дозволяє адаптувати й ті розподіли, для котрих взагалі не було спостережень. В такому процесі всі моделі можуть бути адаптовані й адаптаційний процес динамічно покращується, як тільки з'являється більше адаптаційних даних.

Дерево класів регресії побудовано таким чином, щоб об'єднати компоненти, котрі близькі в акустичному просторі, й, таким чином, схожі компоненти будуть перетворюватися схожим способом. Зазначимо, що дерево побудовано з використанням індивідуальної дикторонезалежної множини моделей фонем, а значить – не залежить від будь-якого нового диктора. Термінальні вузли або листки дерева визначають кінцеві групи компонентів й називаються базовими класами (класами регресії). Кожний гаусіан в множині моделей фонем належить до одного певного базового класу.

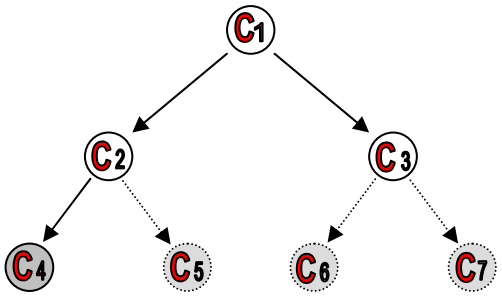

Рис. 1. Бінарне дерево регресії.

На рис.1 наведено простий приклад бінарного дерева регресії з чотирма базовими класами, позначеними як $\{C_4, C_5, C_6, C_7\}$. На діаграмі зображено неперервні стрілки та неперервні околи, й це значить, що адаптаційних даних, пов'язаних з цим класом, достатньо для побудови матриць перетворення. Пунктирні стрілки та околи позначають класи, для яких недостатньо адаптаційних даних. В цьому прикладі вузли 6 та 7 не мають достатньо даних; але у вузлі 3, що є батьківським для 6 та 7, даних достатньо. Аналогічно для вузлів 5 та 2. Кількість даних, що визначається як достатня (поріг), встановлюється як опція вручну.

Перетворення генеруються тільки для тих вузлів, котрі:

1) мають достатньо даних;
2) є або термінальними вузлами (тобто базовими класами), або мають нащадків з недостатньою кількістю даних.

В прикладі, котрий зображений на рис. 1, перетворення генеруються лише для вузлів регресії під номерами 2, 3 та 4, й ці перетворення позначимо відповідно W_2, W_3 та W_4. Звідси, коли потрібно мати перетворену множину моделей фонем, матриці перетворення (для середніх та дисперсій) застосовуються до компонентів гаусіанів в кожному базовому класі наступним чином:

$$\left\{ \begin{array}{c} W_2 \to \{C_5\} \\ W_3 \to \{C_6, C_7\} \\ W_4 \to \{C_4\} \end{array} \right\}$$

Тут цікаво відзначити, що випадок загальної адаптації схожий на випадок, коли дерево має лише один кореневий вузол.

3. Експериментальна база

Як було зазначено у вступі, в даній роботі експерименти проводилися зі спонтанним мовленням. Воно полягає в тому, що диктори, записи котрих використовували в експериментах, говорили вільно, не спеціально для якихось експериментів, порядок слів в їхній мові був вільний, деякі слова вони повторювали й не завжди повністю, говорили з різним ступенем емоційності, в різному темпі, при цьому мовлення було злитим. Розпізнавання також проводилося для злитого мовлення. Каналів запису було багато, вони різнилися між собою за характеристиками. Записи дикторів були не однакового об'єму – від коротких за часом до довгих. Використовувалися записи з теле- та радіоефіру. Всі ці записи були зібрані в так званий корпус АКУЕМ – акустичний корпус українського ефірного мовлення [4]. В цьому корпусі словник налічував 71545 словоформ, біля 60 годин аудіозаписів, в котрих міститься мовлення біля 2000 дикторів. Слід зазначити, що диктори говорили й такі слова, котрих не було в словнику взагалі, на відміну від [1]. Це ускладнювало ситуацію тим, що автоматично понижувало точність розпізнавання. Більшість дикторів представлена короткими записами, тоді як у 150 дикторів довжина записів становить більш як 10 хвилин. Кількість фонем, як й в попередніх дослідженнях, становила 55 елементів. Фонеми моделюються трьома станами Марківського ланцюгу без пропусків. З усього вищесказаного випливає, що, взагалі, умови для розпізнавання в даному випадку менш сприятливі, ніж в попередніх дослідженнях.

4. Результати експериментальних досліджень

Було проведено три експерименти з, відповідно, трьома різними контрольними групами дикторів.

Контрольна група №1 складалася з дикторів, котрі приймали участь в навчанні. Тобто, записи промов цих дикторів були розділені на дві частини: записи з першої частини повністю використовувалися при навчанні системи

розпізнавання (це була навчальна вибірка (НВ)), записи з другої частини використовувалися для тестування та адаптації (це була незалежна вибірка (НезВ) цих дикторів). Мета цього експерименту – експериментально вияснити, коли результати адаптації будуть кращі: коли адаптаційну вибірку (АВ) брати з НВ, чи коли з НезВ? Попутно необхідно було вияснити питання: як залежать результати адаптації від кількості лінійних перетворень, котрі застосовуються при цій самій адаптації? Тобто, кількість адаптаційних даних не змінювалася, АВ залишалася тою самою, а змінювався вручну поріг достатності даних в дереві класів регресії. Чим більший поріг, тим менше буде лінійних перетворень на всю систему при адаптації. Приймалося 4 різних значення порогу – 2000, 1000, 500, 200. Будувалися різні дерева класів регресії – з 1, 2, 3, 4, 6, 8, 10, 13, 16, 20, 25 та 30 термінальними вузлами. Для кожного дерева в залежності від значення порогу обчислювалася різна кількість лінійних перетворень. Результати даного експерименту зображені на рис.2.

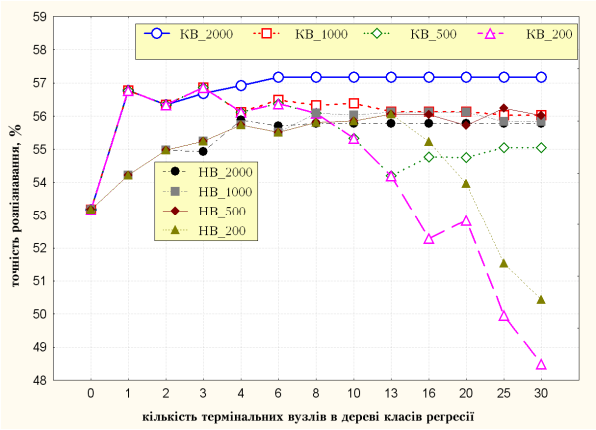

Рис.2. Усереднена точність розпізнавання дикторів з контрольної групи №1 до та після адаптації при 16 гаусіанах в моделях фонем

Пояснення: KB_2000 – це значить, що АВ вибиралася з НезВ, значення порогу 2000; HB_500 – АВ вибиралася з НВ, значення порогу 500. Коли кількість термінальних вузлів – 0, то це значить, що розпізнавання проводилося без адаптації. Досить ясно видно, що результати адаптації кращі, коли АВ вибирають з НезВ (при порогах 2000 та 1000), при порогах 200 та 500 отримуємо досить непевний результат. Виходило, що просте збільшення кількості перетворень (від пониження порогу) без збільшення об'єму АВ не призводить до автоматичного покращення розпізнавання. Можна констатувати, що збільшення точності розпізнавання при виборі АВ з НезВ сягає майже 4% (при порозі 2000), при виборі АВ з НВ сягає майже 3% (при порозі 500). Результати адаптації при виборі АВ з НВ менш розкидані. Дослідження проводилися при кількості гаусіанів в сумішах станів моделей фонем – 16.

Контрольна група №2 складалася з дикторів, котрі не приймали участі в навчанні. Тобто, записи промов цих дикторів не використовувалися при навчанні системи розпізнавання, вони мали лише НезВ. Мета – експериментально вияснити, чи будуть результати адаптації для групи, що не приймала участі в навчанні, кращими, ніж для групи, котра приймала участь в навчанні. Попутно необхідно було вияснити питання: як залежать результати адаптації при збільшенні кількості гаусіанів в сумішах станів моделей фонем? Оскільки в попередньому експерименті при значенні порога 200 отримували незадовільний результат, то тут його не використовували. Дерева класів регресії – ті самі. Результати даного експерименту зображені на рис.3.

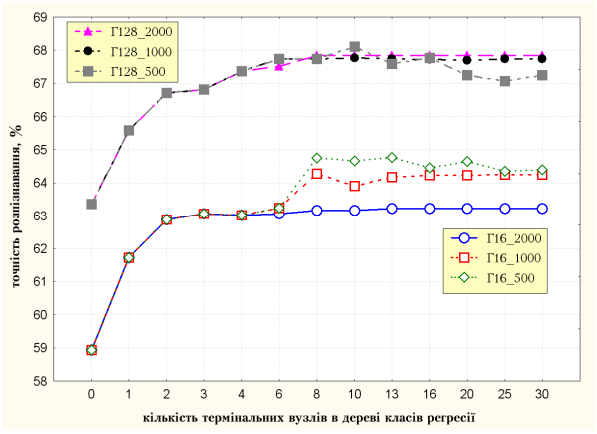

Рис.3. Усереднена точність розпізнавання дикторів з контрольної групи №2 до та після адаптації при 16 та 128 гаусіанах в моделях фонем

Пояснення: Г128_2000 – це значить, що гаусіанів в моделях фонем 128, значення порогу 2000. Чітко видно, що при 128 гаусіанах точність розпізнавання вища як до, так й після адаптації, результати менш розкидані. Зростання точності – до 4,5% (поріг 2000) при 128 гаусіанах, до 5,5% (поріг 500, 1000) при 16 гаусіанах. Порівнюючи з результатами адаптації першого експерименту можна зробити висновок, що при 16 гаусіанах результати адаптації покращилися - 5,5% проти 4%, хоча при цьому говорити про видатну різницю не доводиться.

Контрольна група №3 складалася з дикторів, котрі також не приймали участі в навчанні. Ці диктори – депутати Верховної Ради України, тобто вони говорили зі специфікою парламентських промов й зі специфікою записів цих промов з парламентської зали. Мета – знов таки експериментально вияснити, чи будуть результати адаптації для групи, що не приймала участі в навчанні, кращими, ніж для групи, котра приймала участь в навчанні. Також була поставлена задача: проводити адаптацію не для однієї певної АВ для кожного диктора, а для декількох різних за об'ємом АВ, щоб оцінити якість адаптації в залежності від об'ємів АВ. АВ для всіх дикторів обиралися

об’ємом в 30, 60 та 90 секунд. Дерев класів регресії було побудовано трохи менше. Результати даного експерименту зображені на рис.4.

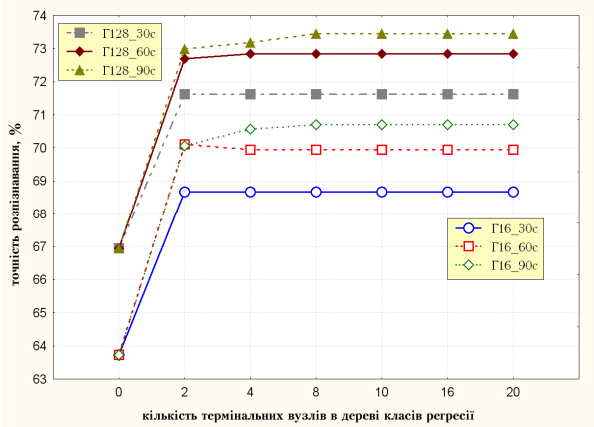

Рис.4. Усереднена точність розпізнавання дикторів з контрольної групи №3 до та після адаптації при 16 та 128 гаусіанах в моделях фонем при значенні порога 2000

Пояснення: Г16_60с – це значить, що гаусіанів в моделях фонем 16, об’єм АВ – 60 секунд. З рисунку видно, що при збільшенні об’єму АВ росте точність розпізнавання після адаптації. Різниця між результатами при АВ в 30 секунд та 60 секунд більша, ніж різниця при АВ в 60 секунд та 90 секунд. При 128 гаусіанах маємо зростання точності після адаптації від 4,5% (при 30с) до 6,5% (при 90с), при 16 гаусіанах - від 5% (при 30с) до 7% (при 90с).

5. Висновки

Отже, експерименти наявно показали доцільність застосування адаптації до голосу нового диктора. Було з’ясовано, що при збільшенні гаусіанів (тут конкретно від 16 до 128) спостерігається покращення точності розпізнавання. Одначе після адаптації більший ріст точності мав місце саме при 16 гаусіанах.

Для дикторів, що приймали участь в навчанні, ріст точності розпізнавання після адаптації був дещо більший тоді, коли АВ вибиралася з НезВ. Для дикторів, що не приймали участі в навчанні, ріст точності розпізнавання після адаптації був дещо більший в порівнянні з дикторами, що приймали участь в навчанні..

Зменшення порогу від 2000 до 200 призводить до збільшення кількості лінійних перетворень для адаптації. Наприклад, коли маємо дерево класів регресії з 30 термінальними вузлами, то кількість перетворень при порозі 2000 – 4, при порозі 1000 – 9, при 500 – 16, при 200 – 25. Цей приклад взятий з конкретного випадку. Для різних випадків (АВ різного об’єму) ці значення будуть дещо різнитися. Експерименти показали, що просте пониження порогу для збільшення кількості перетворень взагалі не призводить до покращення точності. Це стається, очевидно, з причини погіршення статистик внаслідок зменшення кількості спостережень при зменшенні порога.

Експеримент №3 показав, що, взагалі, бажано брати АВ об’ємом не менш за 60 секунд. Хоча й 30 секунд давали зростання точності. При всіх порогах (від 2000 до 200) АВ в 30 секунд завжди давала помітно гірший результат, ніж АВ в 60 секунд та 90 секунд, котрі в свою чергу давали близькі результати. Взагалі, можна зробити той висновок, що збільшення АВ покращує результати адаптації, принаймні до якогось моменту. Задача на подальше – вияснити, коли наступає цей момент, тобто такі об’єми АВ, що подальше нарощування АВ не дає збільшення точності розпізнавання.

Експерименти представили, що ми маємо впевнене зростання точності розпізнавання після адаптації біля 4-5%, хоча в певних варіантах (при АВ в 90с) було й більше. В праці [1] початкове розпізнавання було помітно більшим,. Після адаптації тоді було досягнуто до 6% зростання точності, отже відносне покращення було також суттєво більшим. Але все це відбулося, безсумнівно, внаслідок більш сприятливих умов для розпізнавання.

7. Література

1. Сажок М., Селюх Р., Юхименко О. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови. // Штучний інтелект. – Донецьк, 2009. – № 4. – с. 230-233.

2. М.М.Сажок, Р.А.Селюх, О.А.Юхименко. Адаптація до голосу диктора на основі гендернозалежних акустичних моделей фонем для української мови. – Оброблення сигналів і зображень та розпізнавання образів: Десята Всеукраїнська міжнародна конференція. – Київ, 2010, С.59-62.

3. Young S.J. HTK Book, version 3.1 / Young S.J. [et al]. – Cambridge University, 2002. – 355 p.

4. Н.Б.Васильєва, В.В.Пилипенко, О.М.Радуцький, В.В. Робейко, М.М.Сажок. Створення акустичного корпусу українського ефірного мовлення. – Оброблення сигналів і зображень та розпізнавання образів: Десята Всеукраїнська міжнародна конференція. – Київ, 2010, С.55-58.