Адаптація до голосу диктора на основі гендернозалежних акустичних моделей фонем для української мови

М.М. Сажок, Р.А. Селюх, О.А. Юхименко
Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680
{ mykola, selyukh, yukhymenko } @uasoiro.org.ua

Abstract

Gender dependent systems are usually created by splitting the training data into each gender and building two separate acoustic models for each gender. This method assumes that every state of a subphonetic model is uniformly dependent on the gender. We use the premise that the acoustic realizations of various sub phonetic units are dependent on gender in varying degrees across phones and more particularly context dependent. We show that this is indeed the case by using gender as a question in addition to phone context questions in the context decision trees. Using these trees we build phone specific gender dependent acoustic models and demonstrate a novel method to pick between genders during decoding based on a measure of confidence of the decoded hypothesis. An improvement of 10-20% in word error is achieved relative to a gender independent system.

1. Вступ

Мовленнєвий сигнал містить у собі багато різноманітної інформації, що характеризує особу диктора (мовця), зокрема діалект, функційний стан особи, її вік, стать тощо. Всі ці ознаки є зайвими при розпізнаванні слів. Отже, актуальною є задача позбутися впливу інформації щодо однієї або більше з перелічених ознак, зокрема гендерної ознаки, тобто приналежності диктора до тієї чи іншої статі. При цьому нас цікавить саме пофонемне розпізнавання, оскільки цей підхід дає змогу розширювати словник без навчання на нові слова.

Пофонемне розпізнавання мовленнєвого сигналу передбачає формування усномовного паспорта диктора, що включає акустичні моделі фонем. Оцінка параметрів моделей фонем проводиться за навчальною вибіркою кооперативу дикторів, яка містить фонемне та індивідуальне розмаїття усної мови. Як показують дослідження для англійської мови [], формування акустичних моделей фонем окремо для чоловічої та жіночої статей загалом приводить до підвищення надійності розпізнавання. Втім, слід зауважити, що при такому підході штучно зменшуються обсяги навчальної вибірки, що не є сприятливим для проведення статистичних оцінок.

Для розпізнавання українського мовлення гендерних досліджень взагалі не проводилося.

Авторами поставлено на меті з'ясувати, наскільки впливає стать диктора на розпізнавання окремо вимовлюваних слів у залежності від способів формування усномовного паспорта диктора: або прямо як моделі при розпізнаванні, або як базові моделі (*seeds*) для адаптації на голос диктора.

Зазначимо, що в практичних системах визначення статі диктора може відбуватися автоматично [2,3]. В даній статті ми цього питання не торкаємося.

В наступному розділі описується підхід оцінки параметрів акустичної моделі, розділ 3 присвячено експериментальній базі, в четвертому розділі наводяться результати досліджень, за якими слідують висновки.

2. Постановка задачі адаптації та шляхи її вирішення

Нехай маємо оцінені параметри акустичних генеративних моделей фонем на підставі ітераційних процедур для опорного диктора або для кооперативу дикторів, у який відбираються диктори за гендерною ознакою []. Зокрема, для кожної з трьох фаз-станів фонеми φ (рис. 1) нам відомі вектор математичного сподівання $\mu = [\mu_1, \mu_2, ..., \mu_n]^T$ та коваріаційна матриця Σ, розмірністю $n \times n$, де n – розмірність вектора первинних ознак сигналу.

Рис. 1. Генеративна модель фонеми φ з трьома фазама-станами φ_1, φ_2, φ_3. Додаткові неемітентні стани φ_0 і φ_4 вводяться для сполучення з іншими моделями фонем. Число поруч із дужкою вказує на кількість часових відліків, за які здійснюється перехід.

Припускається, що існує лінійне перетворення, яке переводить початкові вектори математичного сподівання опорного диктора або кооперативу дикторів у вектори математичного сподівання для нового диктора. Це лінійне перетворення представляє собою матрицю розмірністю $n \times n$. Ефектом цього перетворення є зсув середніх значень параметрів моделей фонем та зміна дисперсій цих параметрів у початковій системі таким чином, що кожний стан у системі акустичних моделей фонем буде точніше генерувати дані адаптації.

Лінійне перетворення для середніх значень записується у вигляді:

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

де $\hat{\boldsymbol{\mu}}$ – вектор матсподівання нового диктора, W є матрицею розмірністю $n \times (n+1)$, $\boldsymbol{\xi}$ – вектор розширеного матсподівання,

$$\boldsymbol{\xi} = \left[w, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_n \right]^T , \quad (2)$$

де w представляє нев'язку, початкове значення якої фіксоване і дорівнює 1;

У свою чергу, матриця W розкладається на добуток

$$W = [\, b \; A \,] , \quad (3)$$

де A є матрицею лінійних перетворень розмірністю $n \times n$, а b представляє вектор ухилу.

В такій формі перетворення зручніше обчислюється в умовах неперервного розподілу за нормальним законом.

На відміну від досліджень, представлених у [], у цій роботі розглядалось лінійне перетворення також і коваріаційних матриць, яке записується у вигляді:

$$\hat{\Sigma} = H \; \Sigma H^T , \quad (4)$$

де H – матриця перетворення коваріаційної матриці Σ, розмірність – $n \times n$.

Матриці лінійних перетворень отримуються шляхом оптимізації значення критерію розпізнавання. Одним з таких оптимізаційних алгоритмів є лінійна регресія максимальної правдоподібності (Maximum Likelihood Linear Regression – MLLR) []. Стани фонеми автоматично поділяються на певну кількість класів регресії методами векторного квантування, а потім для кожного класу регресії оцінюється своя матриця лінійних перетворень за ітераційною процедурою.

Ця ж процедура використовується і у випадку апроксимації фаз-станів фонем сумішшю нормальних законів – гаусіанів. Тоді до класів регресії входять окремі гаусіани.

3. Експериментальна база

Були проведені експериментальні дослідження. До експериментів задіяли дикторів з 5 міст України – Коростень (10 дикторів), Кременчук (14 дикторів), Київ (16 дикторів), Львів (14 дикторів), Ніжин (13 дикторів). Всього 67 дикторів (25 чоловіків й 42 диктори жіночої статі). Оскільки є загальновідомим той факт, що надійність розпізнавання жіночих голосів є нижчою [], кількість жінок-дикторів переважає чоловіків. Кожен диктор наговорював свою певну навчальну вибірку (НВ). Оскільки цих певних НВ було 10, то різні диктори могли наговорювати однакові слова. Всього цими дикторами було наговорено 2 416 різних слів. До алфавіту фонем увійшло 55 елементів.

До базового кооперативу дикторів відібрано 53 диктори зі згаданих міст. Решта 14 дикторів увійшли до контрольної групи. Диктори з контрольної групи наговорювали один й той самий набір слів (241 слово). Реалізації цих слів не входять до базового кооперативу.

4. Результати експериментальних досліджень

Результати першого експерименту відображено в Таблиці 1. В ній наведено усереднену надійність розпізнавання до та після адаптації до базового кооперативу дикторів кожного диктора з контрольної групи окремо на 30, 60, 100 та 150 слів.

Кількість гаусіанів у сумішах в моделях фонем – 16.

При адаптації обчислювалися матриці переходу для середнього та дисперсії.

Група дикторів з контрольної групи (14 дикторів) з різних міст, всі наговорювали один й той самий набір слів (241 слово).

Результати, наведені в Таблиці 1, показують, що після адаптації на голос нового диктора надійність розпізнавання в середньому виросла на 3.66% для адаптаційної вибірки обсягом у 30 слів, на 4.45% – для 60 слів, на 5.33% – для 100 слів, на 5.93% – для 150 слів.

Навчання розпізнаванню проводилось на основі бази даних для 53 дикторів з декількох міст України.

При адаптації обчислювалися матриці переходу для середнього та дисперсії.

В Таблиці 2 наведені результати надійності розпізнавання для контрольної групи дикторів жіночої статі до адаптації та після адаптації до кооперативу жінок-дикторів на різну кількість слів. Контрольна група (7 жінок-дикторів) з різних міст, всі наговорювали один й той самий набір слів (241 слово). З таблиці видно, що після адаптації на голос нового диктора надійність розпізнавання в середньому виросла на 2.41% для адаптаційної вибірки обсягом 30 слів, на 2.95% – для 60 слів, на 3.76% – для 100 слів, на 4.46% – для 150 слів. На рис. 2 зображені порівняльні графіки надійності розпізнавання в середньому по контрольній групі жінок-дикторів без врахування гендерності та з врахуванням.

Таблиця 1 – Результати розпізнавання тестових вибірок слів для групи дикторів (14 дикторів) після адаптації на різну кількість слів – 30, 60, 100 та 150 слів.

Кількість слів на адаптацію / Диктори	До адаптації	30	60	100	150
1. Аня	93.78	95.13	95.30	95.32	97.07
2. Анна	91.29	92.76	93.19	93.90	94.51
3. Богдан	80.50	89.71	90.98	92.62	95.24
4. Валентина	95.02	95.26	96.13	96.03	94.87
5. Дмитро	92.12	95.60	96.96	97.73	97.80
6. Катерина	79.25	86.60	87.66	90.21	90.48
7. Олена	90.46	94.11	95.40	95.32	96.34
8. Олеся	92.53	96.82	97.79	98.01	97.80
9. Руслан	89.21	93.23	94.57	95.46	95.24
10. Сергій	95.81	96.41	96.60	97.45	97.80
11. Слава	89.21	93.09	92.81	93.62	93.77
12. Тетяна	87.14	91.33	93.00	94.33	96.33
13. Юрій	89.21	93.16	93.93	96.31	96.70
14. Юрій В.	92.53	96.07	96.04	96.31	97.07
В середньо му по групі	89.86	93.52	94.31	95.19	95.79

Таблиця 2 – Результати розпізнавання тестових вибірок слів для групи жіночих дикторів (7 дикторів) після адаптації на різну кількість слів до кооперативу жінок-дикторів – 30, 60, 100 та 150 слів.

Диктори \ Кількість слів на адаптацію	До адаптації	30	60	100	150
Аня	95.85	96.21	96.60	97.30	98.90
Анна	92.95	93.64	94.02	94.61	96.33
Катерина	84.65	89.37	89.78	92.20	93.04
Олена	93.36	96.07	96.41	96.45	97.44
Олеся	92.95	97.16	97.61	98.15	97.80
Валентина	94.19	94.51	95.58	96.45	95.97
Тетяна	88.80	92.62	93.37	93.90	94.51
В середньому по групі	91.82	94.23	94.77	95.58	96.28
Без врахування гендерності	89.92	93.14	94.07	94.73	95.34

Таблиця 3 – Результати розпізнавання тестових вибірок слів для групи чоловіків-дикторів (7 дикторів) після адаптації на різну кількість слів до кооперативу чоловіків-дикторів – 30, 60, 100 та 150 слів.

Диктори \ Кількість слів на адаптацію	До адаптації	30	60	100	150
Богдан	84.65	89.37	89.96	90.64	92.31
Дмитро	92.95	94.18	95.21	96.31	97.07
Руслан	87.97	94.04	94.67	96.31	95.60
Сергій	96.34	96.55	96.04	98.01	96.70
Слава	91.70	93.57	94.01	94.61	94.87
Юрій	90.04	91.81	93.46	93.48	93.41
Юрій В.	91.29	96.47	95.96	97.02	97.43
В середньому по групі	90.71	93.71	94.19	95.20	95.34
Без врахування гендерності	89.80	93.90	94.56	95.64	96.23

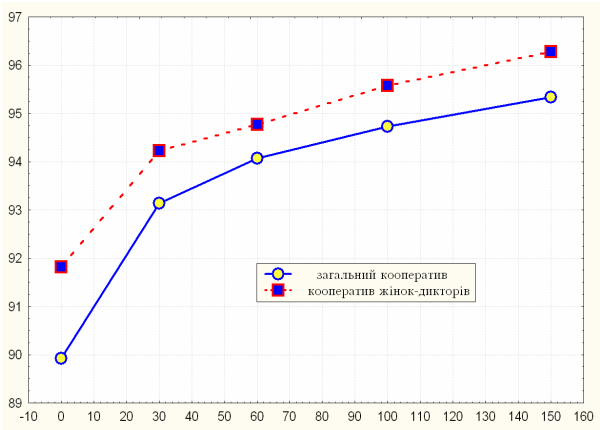

Рис. 2. Усереднена надійність розпізнавання дикторів жіночої статі

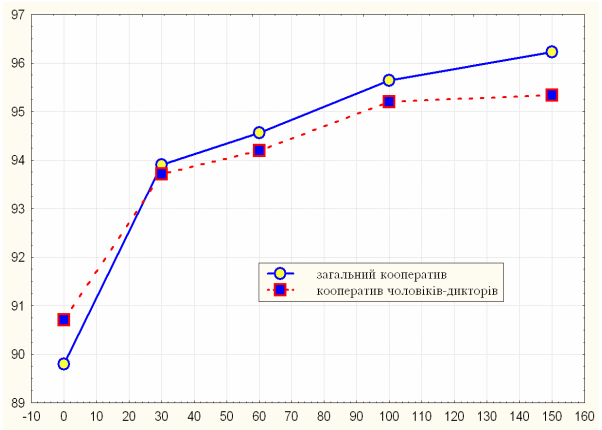

Рис. 3. Усереднена надійність розпізнавання дикторів чоловічої статі

При адаптації обчислювалися матриці переходу для середнього та дисперсії.

В Таблиці 3 наведені результати надійності розпізнавання для контрольної групи дикторів чоловічої статі до адаптації та після адаптації до кооперативу чоловіків-дикторів на різну кількість слів. Контрольна група (7 чоловіків-дикторів) з різних міст, всі наговорювали один й той самий набір слів (241 слово). З таблиці видно, що після адаптації на голос нового

диктора надійність розпізнавання в середньому виросла на 3% для адаптаційної вибірки обсягом 30 слів, на 3.48% – для 60 слів, на 4.49% – для 100 слів, на 4.63% – для 150 слів. На рис. 3 зображені порівняльні графіки надійності розпізнавання в середньому по контрольній групі чоловіків-дикторів без врахування гендерності та з врахуванням.

Навчання розпізнаванню проводилось на основі бази даних для 17 дикторів чоловічої статі з декількох міст України.

При адаптації обчислювалися матриці переходу для середнього та дисперсії.

5. Висновки

Результати гендерозалежного розпізнавання показують зменшення кількості помилково розпізнаних слів до 10–20% порівняно з розпізнаванням на акустичних моделях, сформованих за базовим кооперативом дикторів.

Подальша адаптація до голосу диктора на базі гендернозалежних акустичних моделей показала таку ж динаміку зменшення помилок для дикторів жіночої статі. Цей ефект не спостерігався для чоловічої статі, як ми гадаємо, з причини меншої кількості дикторів-чоловіків у базовому кооперативі.

Подальші роботи будуть спрямовані на підвищення якості адаптації, зокрема шляхом залучення до розпізнавання оцінки довжини голосового тракту диктора. Будуть також досліджені інші простори первинних ознак сигналу. Планується працювати не лише з ізольованими словами, а й зі злитим мовленням, збільшити обсяги словника.

Література

1. Odell, J.J., Woodland, P.C., Valtchev, V., Young, S.J., 1994. Large vocabulary continuous speech recognition using HTK. In: Proceedings of ICASSP, vol. 2. Adelaide, Australia, pp. 125–128.

2. Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов / Т.К. Винцюк. – Киев : Наукова думка, 1987.

3. Peder Olsen and Satya Dharanipragada, "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models," Eurospeech, 4, p. 2509-2512, September 1-4, 2003, Geneva Switzerland.

4. 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.

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