Алгоритм визначення статі диктора за голосом на основі бінарних кодів акустичних подій

Кушнір Д.О., Панфілов С.С.

Група Мовленнєвих Технологій
ТОВ «ДСС Лаб», Москва
d.kushnir@dss-lab.ru, s.panfilov@dss-lab.ru

Abstract

In this paper proposed speaker gender identification algorithm, based on the binarization of acoustic features. The error rate of identification for male is 0.57%, for female - 1.10%.

1. Вступ

На сьогоднішній день найчастіше для вирішення задачі визначення статі диктора використовуються підходи на базі моделей гауссових сумішей (GMM - Gaussian Mixture Model) в різних ознакових просторах (MFCC, LPC, PLP). Такі моделі можуть бути навчені як шляхом незалежного навчання на вибірці кожного класу, так і шляхом адаптації фонової моделі (UBM - Universal Background Model). У разі використання GMM помилка класифікації досягає рівня не більше 1-2% для кожного класу [1,2]. Не менш якісний результат дає підхід, заснований на моделюванні мовного тракту, розроблений Сорокіним [3].

Нашою лабораторією запропонований простий, швидкий і ефективний підхід, що використовує UBM-GMM, побудованої в архітектурі бінарного дерева, яка дозволяє кодувати акустичні події (вектор ознак) в цілочисельний код. В процесі побудови моделі прийняття рішень відбувається накопичення статистик по кожному коду акустичних подій. В процесі ідентифікації статі диктора, накопичені статистики використовуються для обчислення оцінки гендерної приналежності.

2. Опис підходу

2.1. Корпус

При розробці системи були використані три багато-мовних корпуса

- для побудови UBM (200 годин).
- для побудови гендерной модели (10 годин приблизно в рівних частках для чоловіків і жінок).
- для тестування (90 годин чоловічих і жіночих голосів у співвідношенні 1:2).

2.2. Акустичні ознаки

В якості базових ознак використовуються мел-спектральні кепстральні коефіцієнти (MFCC - mel-frequency cepstral coefficients). В частотній смузі 50-3600 Гц задається гребінка з 17 мел-фільтрів, з яких для обчислення MFCC беруться останні 16.

2.3. Універсальна фонова модель

UBM – це GMM, побудована на великій кількості даних. В межах завдань ідентифікації необхідно максимально повно представити дикторський базис. Для навчання UBM ми використовували бази дикторських голосів російською, англійською, французькою, німецькою, італійською, арабською, китайською та японською мовою. Використовувана кількість гауссіан в суміші - 1024.

Структурно UBM виконана у вигляді бінарного дерева. Такий спосіб організації зберігання гауссіан забезпечує швидкий пошук найближчих компонент суміші, складність пошуку складає $O(\log_2 N)$.

Варто зауважити, що основну обчислювальну складність алгоритму ідентифікації становить обчислення набору акустичних ознак, а точніше обчислення швидкого перетворення Фур'є: $O(N\log_2 N)$, N - розмір вікна аналізу.

2.4. Моделювання жіночих та чоловічих голосів

2.4.1. Обчислення бінарних кодів акустичних подій

Для кожного вектора акустичних ознак FV обчислюється відповідний бінарний код акустичної події (КАП), $\mathrm{BinFV} = [0|1]$. Код визначається траєкторією обходу дерева UBM, яка виходить у процесі класифікації вектора FV, а також положенням вектора ознак щодо центру найближчій до нього гауссіана за напрямками головних осей. При цьому вектор ознак кодується одиницею, якщо він знаходиться на позитивній півосі і нулем, якщо на негативній. Розрядність коду акустичного події визначається кількістю рівнів GMM і кількістю головних напрямків найближчій гауссіана, які використовуються при кодуванні. У даний роботі 10 біт коду кодували номер гауссіана і додаткові 10 біт кодували позицію вектора ознак щодо центру найближчій гауссіана. Таким чином, кожне акустичне подія кодувалося 20-ти бітовим цілочисловим значенням. Варто відзначити, що такий спосіб кодування акустичних подій здійснюється в порядку зменшення значимості біта, що кодує. Кодування положення вектора ознак в контексті всієї суміші гауссіан

кодується старшими бітами, причому, чим старше біт, тим більша частина простору ознак враховується при кодуванні, а положення вектора ознак в околиці однієї гауссіана за напрямками вздовж головних осей кодується молодшими бітами також в порядку убування власних чисел коваріаційної матриці.

2.4.2. Гендерна модель

Моделювання статі диктора здійснювалося за допомогою накопичення статистик кодів акустичних подій. Статистикою КАП в роботі зветься зустрічальність відповідної акустичної події у чоловіків і жінок. У нашій роботі передбачається, що більшість КАП мають виражену гендерну належність, що повинно відбиватися в накопичених статистиках.

На навчальних множинах M_female і M_male обчислюється хеш-таблиця *H*.

$$H=\{c_n;(m_n,f_n)\},\quad n=1..N,\quad N=|КАП^M \cup КАП^F| \tag{1}$$

c_n – ключ, цілочисельне значення, отримане за бінарним кодом; m_n – частота коду у вибірці чоловіків; f_n – частота коду у вибірці жінок; $|КАП|^M$ – множина кодів, які були зустрінуті в вибірці чоловіків; $|КАП|^F$ – множина кодів, які були зустрінуті в вибірці жінок.

З метою зниження «шумового ефекту» в статистиках, а також для використання найбільш значущих статистик, проводилась процедура відсіву малозначущих статистик КАП. Для цього обчислюються математичне сподівання μ_S і значення середньоквадратичного відхилення σ_S «жіночих» (S=f) та «чоловічих» (S=m) статистик. З хеш-таблиці виключаються такі статистики S, значення яких виходять за інтервал $[\mu_S-3\sigma_S;\mu_S+3\sigma_S]$. Така процедура суттєво скорочує число використовуваних КАП і, отже, зменшується розмір хеш-таблиці без істотної втрати в якості класифікації.

При обчисленні оцінок приналежності КАП тієї чи іншої статі застосовувалася процедури нормалізації. На першому етапі значення статистик наводяться до інтервалу [0; 1]:

$$S_n^{'}=\frac{S_n}{m_n+f_n},\quad S=\{m|f\} \tag{2}$$

Це дає оцінку характерності даного КАП для чоловіків і жінок. На другому етапі проводиться нормалізація до обсягів вибірок по чоловікам і жінкам:

$$S_n^{''}=\frac{S_n^{'}}{\sum_{n=1}^{N} S_n^{'}},\quad S=\{m|f\} \tag{3}$$

2.5. Ідентифікація

Алгоритм:

- Обчислення послідовності КАП

$$Y=(y_1,y_2,\ldots,y_T) \tag{4}$$

- Підрахунок оцінок

$$M\_score=\sum_{t=1}^{T}\left(H(y_t).m:y_t\in H\right) \tag{5}$$

$$F\_score=\sum_{t=1}^{T}\left(H(y_t).f:y_t\in H\right) \tag{6}$$

- Прийняття рішення за принципом максимуму

$$\max(M\_score,F\_score) \tag{7}$$

3. Тестування

Результати тестування наведені у таблиці 1.

Таблиця 1: Результати тестування

Голоси	Помилка
Чоловічі	0,57%
Жіночі	1,10%

4. Висновки

Запропонований алгоритм показав добрі результати класифікації: для чоловічих голосів - 99,43%, для жіночих - 98,9%.

У подальші плани входить розробка таких алгоритмів бінаризації акустичних подій та їх класифікації, які максимально не корелюють із запропонованим алгоритмом і при об'єднанні з ним поліпшать результат в цілому.

5. Література

[1] Калюжный А. Я., Семёнов В. Ю. "Автоматическое определение пола диктора на основе гауссовых смесей", *Акустический симпозиум «Консонанс-2009»: сборник тезисов конференции*. 31 с, 2009.

[2] Yücesoy E., Nabiyev V. V. "Gender identification of a speaker using MFCC and GMM", *IEEE Electrical and Electronics Engineering (ELECO):626-629, 1999*

[3] Сорокин В. Н., Тананыкин А. А. "Распознавание пола диктора с помощью метода Парзена", *Журнал "Доклады Томского государственного университета систем управления и радиоэлектроники"*, № 4(30):159-162, 2013