КОМПЬЮТЕРНАЯ
МАТЕМАТИКА

УДК 004.934.2

В.Ю. СЕМЕНОВ

МЕТОД КЛАССИФИКАЦИИ ПОЛА ДИКТОРА, ОСНОВАННЫЙ НА АППРОКСИМАЦИИ ПАРАМЕТРОВ ГОЛОСА МОДЕЛЬЮ ГАУССОВЫХ СМЕСЕЙ

Предложен метод классификации пола диктора, основанный на моделировании плотности распределения вектора параметров голоса моделью гауссовых смесей. В качестве вектора параметров использовался набор кепстральных коэффициентов, дополненный периодом основного тона. Обучение модели гауссовых смесей производилось по методу максимизации ожидания (Expectation-Maximization). При проведении экспериментальных исследований предложенный метод показал достаточно низкую вероятность ошибки классификации (от 9 % до 0 %).

Вступление. Задача идентификации пола диктора актуальна для систем автоматической классификации речевой информации, поскольку предварительное определение пола обеспечивает более точную настройку распознающей системы. Кроме того, определение пола диктора может представлять самостоятельный интерес в системах, обеспечивающих правоохранительную деятельность, сбор информации для рекламных целей и т. п. [1]. Как известно, ключевыми вопросами для построения любой системы распознавания являются:

• выбор признаков, т. е. параметров, характеризующих распознаваемые объекты (в данном случае – мужские/женские голоса);

• выбор модели, в соответствии с которой производится обучение системы распознавания и последующая классификация признаков.

Как правило, на предварительном этапе из базы тестовых сигналов выделяются вектора (наборы) признаков. Затем полученный массив признаков используется для обучения классифицирующей модели, результатом чего являются некоторые классы или эталонные значения признаков.

В процессе реальной работы проверяемый сигнал подвергается предварительной обработке (масштабирование, удаление шумов). Затем извлеченные из него признаки сравниваются с полученными на предварительном этапе эталонными значениями (решающее правило), что и является результатом классификации.

В задачах распознавания речевой информации в роли вектора признаков обычно выступают кепстральные параметры, вычисляемые на каждом фрейме речевого сигнала. В качестве вектора параметров нами выбран набор из десяти RASTA-PLP коэффициентов, дополненный периодом основного тона (ОТ). При этом был исключен кепстральный коэффициент, отвечающий за уровень сигнала, т. е. общая размерность вектора признаков составляла 11.

В задачах распознавания используются различные подходы к выбору модели классификации: Модели Гауссовых Смесей (Gaussian Mixture Models, GMM), Скрытые Марковские Модели (Hidden Markovian Models, HMM) и другие. Выбор между методами GMM и HMM зависит от того, является ли поставленная задача идентификации текстонезависимой или текстозависимой. В данном случае целесообразно принять, что обсуждаемая задача является текстонезависимой, поскольку нас интересует не динамическая смена признаков, а интегральное преобладание одних над другими. С учетом данного обстоятельства выбран аппарат GMM.

Вычисление признаков. Важным признаком, используемым для различия мужских и женских голосов, является период основного тона T_0 (или частота основного тона (ОТ) $f_0 = 1/T_0$). Этот параметр характеризует частоту колебания голосовых связок при произнесении звонких звуков [2].

Как правило, для мужчин характерны меньшие значения частоты основного тона по сравнению с женщинами. Однако, как видно из рис. 1, эти диапазоны все же пересекаются, так что в некоторых случаях женскому голосу может соответствовать меньшая частота ОТ. Поэтому наиболее сложными представляются ситуации, когда необходимо идентифицировать женщину с низким голосом или, наоборот, мужчину с высоким голосом. В таких случаях правильная идентификация должна обеспечиваться за счет использования параметров, отражающих различия в структуре голосового тракта мужчин и женщин. Исходя из вышесказанного, мы включили в вектор признаков

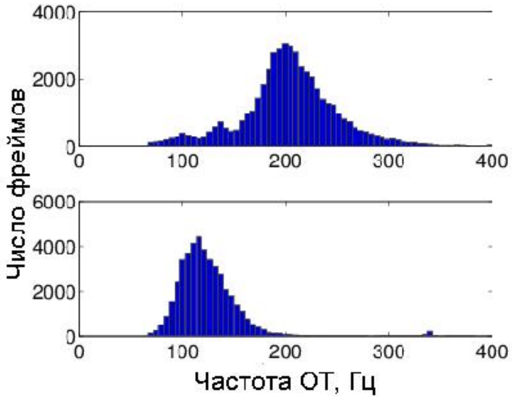

РИС. 1. Гистограммы распределения частоты основного тона (Pitch) для женщин (верхний график) и для мужчин (нижний график)

10 RASTA-PLP коэффициентов, определяющих форму голосового тракта при произнесении звуков. Методика анализа речевых сигналов RASTA-PLP состоит из двух частей: PLP (Perceptual linear prediction [3]) – линейное предсказание с учетом особенностей слухового восприятия, RASTA («RelAtive SpecTrA») – обработки, предназначенной для удаления из сигнала спектральных компонент, скорость изменения которых отлична от скорости изменения соответствующих компонент речи [4]. Основные этапы этой обработки перечислены далее.

1. *Разбивка на фреймы***.** При работе с частотой дискретизации $f_s = 8000$ Гц использовались фреймы длиной 25 мс (200 дискретных отсчетов) с перекрытием в 15 мс (120 дискретных отсчетов).

2. *Вычисление спектра.* На каждом фрейме вычисляется квадрат модуля преобразования Фурье.

3. *PLP-анализ.* Частотный диапазон $[0, f_s/2]$ разбивается на 17 критических полос [3], в каждой из которых производится вычисление энергии голосового сигнала E_i, $i = 1,...,17$.

4. *RASTA-фильтрация***.** Сглаживание логарифмов величин E_i с помощью цифрового фильтра [4] с дискретной передаточной функцией

$$R(z) = 0.1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}}. \tag{4}$$

Через фильтр $R(z)$ пропускается каждая из 17 спектральных траекторий $\log E_i$, полученных на предыдущем этапе. На рис. 2 показаны траектории логарифмов энергий в критических полосах до и после RASTA-фильтрации. Сравнение показывает, что RASTA-фильтрация убирает постоянные составляющие логарифмов спектральных компонент.

5. *Умножение на кривую равной громкости.* Сглаженный логарифмический спектр, полученный в результате RASTA-фильтрации, возвращается в линейный масштаб путем взятия от него экспоненты.

Затем на каждом фрейме спектр умножается на кривую равной громкости [3], которая определяется соотношением

$$H(f) = \frac{f^4}{(f^2 + 1.6 \times 10^5)^2} \times \frac{f^2 + 1.44 \times 10^6}{f^2 + 9.61 \times 10^6}, \tag{5}$$

где f – частота в линейном масштабе. График данной функции в логарифмическом масштабе показан на рис. 3.

6. *Имитация закона слухового восприятия.* Полученные на предыдущем шаге спектры для каждого фрейма возводятся в степень 0.33.

7. *Обратное преобразование Фурье.* От спектра берется обратное преобразование Фурье, результатом чего является автокорреляционная функция (АКФ) $R(k)$, $k = 0,\dots,L_{fft} - 1$.

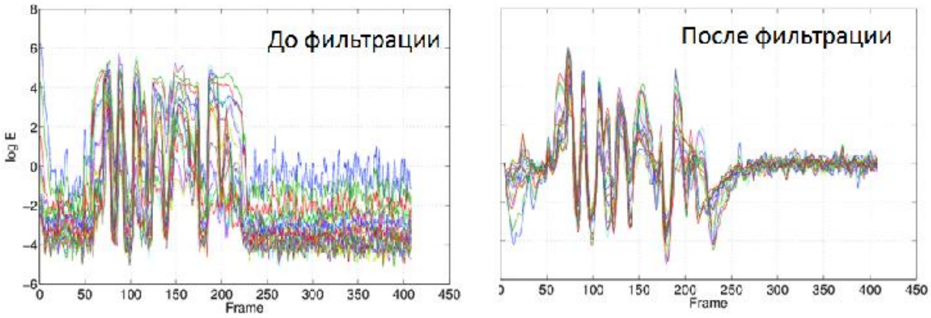

РИС. 2. Пример траекторий логарифмов энергий в критических полосах до и после RASTA-фильтрации

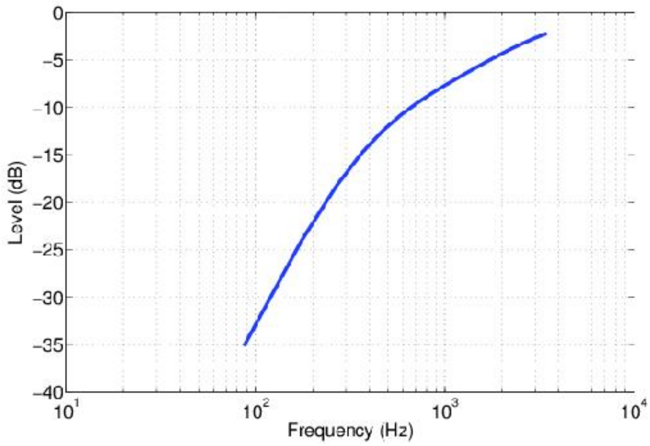

РИС. 3. Кривая равной громкости

8. *Вычисление коэффициентов линейного предсказания* (*КЛП*). Для вычисления КЛП $a_1, a_2, ..., a_p$ (в нашем случае $p = 10$), нам необходимы первые $(p+1)$ значений АКФ: $R(0), R(1), \ldots, R(p)$. Они вычисляются с помощью рекурсии Левинсона – Дарбина [2].

9. *Преобразование в кепстральные коэффициенты.* Кепстральные коэффициенты (т. е. обратное преобразование Фурье от логарифма АР спектра) вычисляются через рекуррентные соотношения [2]:

$$c_n = -a_n - \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, n = 1, ..., p.$$

На завершающем этапе коэффициенты проходят процедуру «лифтинга»:

$$c_{n'} = n^{0.6} c_n,\ n = 1, \ldots,\ p.$$

Таким образом, итоговый вектор признаков состоит из периода основного тона T_0 и RASTA-PLP коэффициентов $c_1', c_2', ..., c_{10}'$.

Модель гауссовых смесей (GMM). Основная идея аппарата GMM состоит, как известно [5], в представлении плотности распределения вектора акустических параметров x (размерностью d) в виде взвешенной суммы гауссовских плотностей распределения:

$$p(x) = \sum_{m=1}^{M} \alpha_m b(x / \mu_m, D_m), \tag{6}$$

где $b(x / \mu, D)$ – гауссова плотность со средним μ и ковариационной матрицей D:

$$b(x / \mu, D) = \frac{1}{\sqrt{2\pi \det D}} \exp(-0.5(x - \mu)^T D^{-1}(x - \mu)). \tag{7}$$

Фактически представление плотности $p(x)$ в виде суммы M гауссианов соответствует разбиению множества акустических параметров на M подклассов [5]. Такой подход схож с идеей векторного квантования, но является при этом более гибким.

Заметим, что для GMM не важен порядок следования друг за другом акустических единиц (фонем и др.), этот аппарат работает с накопленными статистиками параметров.

GMM должны быть независимо обучены для каждого из альтернативных классов дикторов (т. е. для мужского и женского). Это означает, что для каждого класса должен быть найден свой набор параметров $\alpha_i, \mu_i, D_i, i = 1, ..., M$. Исходные данные для обучения – это набор векторов акустических признаков $X = x_1, x_2, ..., x_T$.

Обучение GMM традиционно осуществляется по алгоритму EM (expectation-maximization) [6]. Существуют два варианта для вычисления ковариационных матриц D_i, предполагающие их «полную» или диагональную структуру [5].

Приведем уравнения для итеративного вычисления параметров $\alpha_i, \mu_i, \sigma_i, i = 1, ..., M$ в случае диагональных ковариационных матриц [5]. Обновление апостериорных вероятностей попадания в m-й класс выполняется по формуле:

$$p(m / x_i, \alpha, \mu, \sigma) = \frac{\alpha_m b_m(x_i)}{\sum_{m=1}^{M} \alpha_m b_m(x_i)},$$

где

$$b_m(x_i) = \frac{\exp\{-0.5 \sum_{k=1}^{d} (x_i^k - \mu_m^k)^2 / (\sigma_m^k)^2\}}{\prod_{k=1}^{d} \sigma_m^k}.$$

Обновление весов:

$$\alpha_m = \frac{1}{N}\sum_{i=1}^{N} p(m / x_i, \alpha, \mu, \sigma);$$

обновление средних значений:

$$\mu_m = \frac{\sum_{i=1}^{N} p(m / x_i, \alpha, \mu, \sigma) x_i}{\sum_{i=1}^{N} p(m / x_i, \alpha, \mu, \sigma)};$$

обновление дисперсий:

$$(\sigma_m)^2 = \frac{\sum_{i=1}^{N} p(m / x_i, \alpha, \mu, \sigma)(x_i)^2}{\sum_{i=1}^{N} p(m / x_i, \alpha, \mu, \sigma)} - (\mu_m)^2.$$

В работе [7], рекомендуется использовать 15 итераций алгоритма EM, а в более поздней работе того же автора [8], – 5 итераций.

Как обычно, для метода EM остро стоит проблема начальной инициализации. Хотя этому вопросу уделено большое внимание в литературе, в работе [5] делается замечание, что в конечном итоге результаты идентификации диктора не сильно зависят от способа инициализации алгоритма в процессе тренинга GMM. В данной работе для инициализации обучения GMM мы использовали алгоритм K-средних [9].

Применение алгоритма K-средних к набору векторов акустических признаков $X = x_1, x_2, ..., x_T$ позволяет найти M квантов, служащих инициализацией для математических ожиданий $\mu_m, m = 1, ..., M$. Далее, отбирая вектора x_i попавшие в m-ю ячейку $K^{(m)}$, мы получаем приближение для дисперсий

$$(\sigma_m^k)^2 = \frac{\sum_{i \in K^{(m)}} (x_i^k - \mu_m^k)^2}{N^{(m)}}, \; k = 1, ..., d,$$

где $N^{(m)}$ – количество элементов в m-й ячейке.

Значения α инициализируются как

$$\alpha_m = \frac{N^{(m)}}{N}.$$

В процессе реальной работы, имея набор из N наблюдений $X = x_1, x_2, ..., x_N$, проверка гипотез сводится к простому сравнению плотностей вероятностей $p(X / \alpha^{(male)}, \mu^{(male)}, D^{(male)})$ и $p(X / \alpha^{(female)}, \mu^{(female)}, D^{(female)})$. Предполагая независимость векторов наблюдений, эти величины удобно записывать в нормированном логарифмическом масштабе:

$$L^{(male)} = \frac{1}{N}\log p(X / \alpha^{(male)}, \mu^{(male)}, D^{(male)}) = \frac{1}{N}\sum_{i=1}^{N}\log p(x_i / \alpha^{(male)}, \mu^{(male)}, D^{(male)}),$$

$$L^{(female)} = \frac{1}{N}\log p(X / \alpha^{(female)}, \mu^{(female)}, D^{(female)}) =$$

$$= \frac{1}{N}\sum_{i=1}^{N}\log p(x_i / \alpha^{(female)}, \mu^{(female)}, D^{(female)}),$$

где $\log p(x_i / \alpha^{(male)}, \mu^{(male)}, D^{(male)})$ и $\log p(x_i / \alpha^{(female)}, \mu^{(female)}, D^{(female)})$ записываются в соответствии с (6). Если $L^{(male)} > L^{(female)}$, выносится решение о преобладании мужского голоса. В противном случае, выносится решение о преобладании женского голоса.

Экспериментальные результаты. Для проверки метода сформированы две базы речевых сигналов, описываемые далее.

База 1. В формировании записей участвовали 16 мужчин и 11 женщин. Среди языков были представлены русский и английский (США). Для каждого диктора взято 10 файлов. Общая их продолжительность составила около 8 минут для мужчин и около 6 минут для женщин.

База 2. В формировании записей участвовали 21 мужчина и 13 женщин. Среди языков были представлены португальский (Бразилия), английский, немецкий, хинди, венгерский, японский, русский, испанский. Общая их продолжительность составила около 20 минут для мужчин (103 файла) и около 20 минут для женщин (154 файла).

В первом эксперименте база 1 была взята в качестве обучающей, а база 2 – в качестве проверочной. Во втором эксперименте база 2 была взята в качестве обучающей, а база 1 – в качестве проверочной. При этом количество компонент GMM взято равным 1, 2, 4, 8, 12, 16.

В табл. 1 и 2 приведены проценты ошибок классификации при различных порядках GMM, а также разных типах ковариационных матриц.

Более низкий процент ошибок при использовании базы 2 в качестве обучающей, объясняется ее большим объемом (по сравнению с базой 1) и разнообразием дикторов базы 2. Кроме того, как видно из таблиц, увеличение числа компонент GMM не приводит к уменьшению ошибки распознавания. Этот результат подтверждается работой [10], где модификация с двумя гауссианами обеспечивала практически такой же процент ошибок, как и для $M = 4, 6, 8, 10, 12, 16$.

Подводя итоги, можно сказать, что порядок и тип GMM не влияют существенно на процент ошибок при классификации голосов по половому признаку. Главным фактором является разнообразие дикторов в обучающей базе речевых сигналов.

ТАБЛИЦА 1. Процент ошибок при использовании базы 1 в качестве обучающей, а базы 2 – в качестве проверочной, для диагональных и полных ковариационных матриц различных размерностей

	Диаго-нальная 1	Диаго-нальная 2	Диаго-нальная 4	Диаго-нальная 8	Диаго-нальная 12	Диаго-нальная 16
Мужчины	8.7 %	4.9 %	3.9 %	1.0 %	3.9 %	3.9 %
Женщины	9.1 %	7.1 %	7.8 %	7.8 %	7.1 %	7.1 %
Среднее	8.9 %	6.2 %	6.2 %	5.1 %	5.8 %	5.8 %
	Полная 1	Полная 2	Полная 4	Полная 8	Полная 12	Полная 16
Мужчины	3.9 %	3.9 %	1.9 %	3.9 %	1.9 %	2.9 %
Женщины	4.5 %	6.5 %	6.5 %	7.1 %	7.1 %	7.1 %
Среднее	4.3 %	5.4 %	4.7 %	5.8 %	5.1 %	5.4 %

ТАБЛИЦА 2. Процент ошибок при использовании базы 2 в качестве обучающей, а базы 1 – в качестве проверочной,для диагональных и полных ковариационных матриц различных размерностей

	Диаго-нальная 1	Диаго-нальная 2	Диаго-нальная 4	Диаго-нальная 8	Диаго-нальная 12	Диаго-нальная 16
Мужчины	0.7 %	0.7 %	0.0 %	0.0 %	0.0 %	0.0 %
Женщины	0.9 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Среднее	0.8 %	0.4 %	0.0 %	0.0 %	0.0 %	0.0 %
	Полная 1	Полная 2	Полная 4	Полная 8	Полная 12	Полная 16
Мужчины	0.7 %	0.7 %	0.7 %	0.7 %	0.7 %	0.7 %
Женщины	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Среднее	0.4 %	0.4 %	0.4 %	0.4 %	0.4 %	0.4 %

Наиболее выгодными с практической точки зрения видятся модификации с диагональными ковариационными матрицами размером 4×4 или 8×8, поскольку они дают на общем фоне приемлемый процент распознавания и при этом характеризуется существенно меньшими вычислительными затратами по сравнению с использованием полных матриц.

Выводы. В данной работе предложен автоматический классификатор пола диктора на основе моделирования акустических параметров голоса с помощью аппарата гауссовых смесей (GMM). В качестве вектора акустических признаков выбран вектор кепстральных RASTA-PLP коэффициентов, дополненный периодом основного тона.

Результаты испытаний показывают процент ошибок классификации от 9 % до 0 % в зависимости от объема обучающей и проверочной баз, типа ковариационных матриц GMM (полные/диагональные) и их порядков.

Порядок и тип GMM являются вторичным фактором для правильной классификации пола диктора по сравнению с необходимостью разнообразия дикторов в обучающей базе речевых сигналов.

Наиболее практичными видятся модификации с диагональными ковариационными матрицами малого размера (например, 4×4 или 8×8), поскольку они дают на общем фоне приемлемый процент распознавания и при этом характеризуется существенно меньшими вычислительными затратами по сравнению с использованием полных матриц.

В.Ю. Семенов

МЕТОД КЛАСИФИКАЦІЇ СТАТІ ДИКТОРА, ЗАСНОВАНИЙ НА АПРОКСИМАЦІЇ ПАРАМЕТРІВ МОВНОГО СИГНАЛУ МОДЕЛЛЮ ГАУССІВСЬКИХ СУМІШЕЙ

Запропоновано метод класифікації статі диктора, заснований на моделюванні щільності розподілу вектора параметрів мовного сигналу моделлю гауссівських сумішей. Як вектор параметрів використовувався набір кепстральних коефіцієнтів, доповнений періодом основного тону. Навчання моделі гауссівських сумішей здійснювалось за методом максимизації очікування (Expectation-Maximization). При експериментальній перевірці запропонований метод показав досить низьку ймовірність помилки класифікації (від 9 % до 0 %).

V. Semenov

METHOD FOR GENDER CASSIFICATION BASED ON APPROXIMATION OF VOICE PARAMETERS BY GAUSSIAN MIXTURE MODELS

The method for gender classification, based on modeling of probability density function of voice parameters by Gaussian mixture model, is proposed. The vector of parameters consists of cepstral coefficients combined with basic tone period. The training of Gaussian mixture model is performed by Expectation-Maximization method. The experiments have shown low probability of classification error for the proposed method (from 9% to 0%).

Список литературы

1. Li M., Han K. J., Narayanan S. Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level In formation Fusion. *Computer Speech and Language*. 2013, Vol. 27. P. 151–167.
2. Рабинер Л., Шафер Р. Цифровая обработка речевых сигналов. М.: Радио и связь, 1981. 496 с.
3. Hermansky H., Morgan N. RASTA processing of speech. *IEEE Trans. Speech and Audio Processing*. 1994. Vol. 2, N 6. P. 578–589.
4. Hermansky H. Perceptual Linear Prediction (PLP) analysis of speech. *J. Acoust. Soc. America*. 1990. Vol. 87. P. 1738–1753.
5. Reynolds D.A., Rose R.C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Trans. Speech Audio Proces.* 1995. Vol. 3. P. 72–83.

6. Dempster A., Lair N., Rubin D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statistical Society*. 1977. **39**. P. 1 – 38.
7. Reynolds D.A. Experimental Evaluation of Features for Robust Speaker Identification. *IEEE Trans. Speech Audio Proces.* 1994. Vol. 2. P. 639–643.
8. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*. 2000. Vol. 10. P. 19 – 41.
9. Linde Y., Buzo A., Gray R.M. An Algorithm for Vector Quantizer Design. *IEEE Trans. Communications.* 1980. Vol. 28, N. 1. P. 84 – 95.
10. Zeng Y.-M., Wu Z.-Y., Falk T., Chang W.-Y. Robust GMM-based gender classification using pitch and RASTA-PLP parameters of speech. *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics.* 2006. P. 3376–3379.

Получено 17.09.2018

Об авторе:

Семенов Василий Юрьевич,
кандидат физико-математических наук,
докторант Института кибернетики имени В.М. Глушкова НАН Украины.
E-mail: vasyl.delta@gmail.com