АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ПОЛА ДИКТОРА НА ОСНОВЕ ГАУССОВЫХ СМЕСЕЙ

А. Я. КАЛЮЖНЫЙ, В. Ю. СЕМЕНОВ

ГНПП "Дельта", Киев
e-mail: {kalyuzhny,semenov}@deltacorp.net

Предложен метод автоматической классификации речевых фрагментов по признаку "мужчина/женщина". Метод основан на моделировании плотности распределения вектора акустических признаков голоса взвешенной суммой нескольких гауссовских распределений (метод гауссовых смесей, GMM [5]). При этом каждый член GMM соответствует некоторому подклассу множества акустических параметров голосового сигнала. В качестве вектора акустических признаков был выбран вектор кепстральных коэффициентов (PLP), дополненный периодом основного тона. Для повышения помехоустойчивости метода при вычислении PLP коэффициентов выполнялась RASTA-фильтрация [1]. Обучение гауссовых смесей производилось на речевой базе, включающей 8 иностранных языков. Вычисление параметров гауссовых смесей для мужских и женских голосов производилось по методу Expectation-Maximization с инициализацией согласно алгоритму K-средних. Результаты предварительных испытаний на базе русскоговорящих дикторов, не принимавших участие в формировании тестовой базы, свидетельствуют об устойчивой работе метода. Вероятность правильного распознавания пола диктора при этом составляла от 98 до 100% для различных модификаций предложенного метода.

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 рекламной информации и т.п.

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.

В роли вектора признаков обычно выступает вектор из 10-20 кепстральных параметров, вычисляемых на каждом фрейме речевого сигнала. По аналогии с работой [3], нами в качестве вектора параметров был выбран набор RASTA-PLP коэффициентов, дополненный периодом основного тона (ОТ). Это соответствует 12-мерному вектору параметров (1 - основной тон и 11 коэффициентов RASTA-PLP). При этом, согласно рекомендациям литературы по идентификации диктора, нами был исключен кепстральный коэффициент, отвечающий за уровень сигнала, т.е. использовался 11-мерный вектор параметров.

В задачах распознавания используются различные подходы к классификации: Гауссовы Смеси (Gaussian Mixture Models, GMM), Скрытые Марковские Модели (HMM),

Support Vector Machine (SVM) и другие. Выбор между подходами GMM и HMM зависит от того, является ли поставленная задача идентификации текстонезависимой или текстозависимой. Применительно к обсуждаемой задаче, можно принять, что она является текстонезависимой, поскольку нас интересует не динамическая смена признаков, а интегральное преобладание одних над другими. С учетом данного обстоятельства был выбран аппарат GMM.

1 ВЫЧИСЛЕНИЕ ПРИЗНАКОВ

1.1 Период основного тона

Важным признаком, используемым для различения мужских и женских голосов, является период основного тона T_0 (или частота основного тона $f_0 = 1/T_0$). Этот параметр характеризует частоту колебания голосовых связок при произнесении звонких звуков. Для вычисления периода основного тона нами использовался автокорреляционный метод, описанный в работе [4].

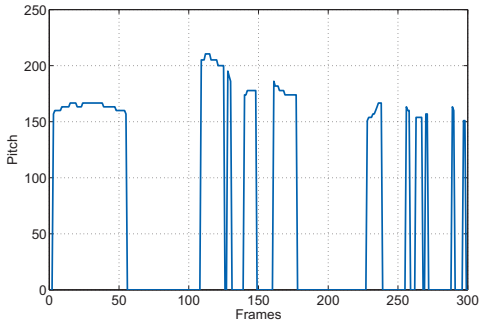

Рис. 1. Пример изменения частоты основного тона (pitch) для англоязычного диктора-мужчины

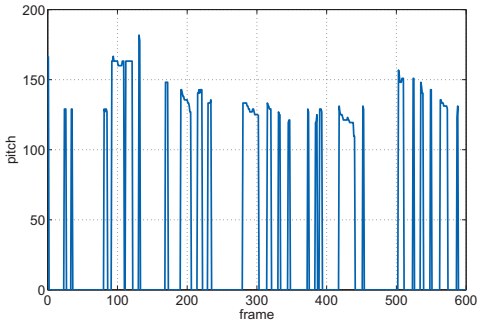

Рис. 2. Пример изменения частоты основного тона (pitch) для японоязычного диктора-женщины

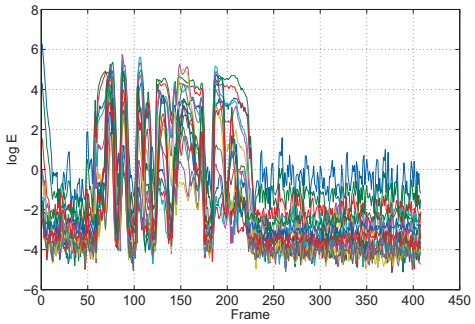

Рис. 3. Пример траекторий логарифмов энергий в критических полосах до RASTA-фильтрации

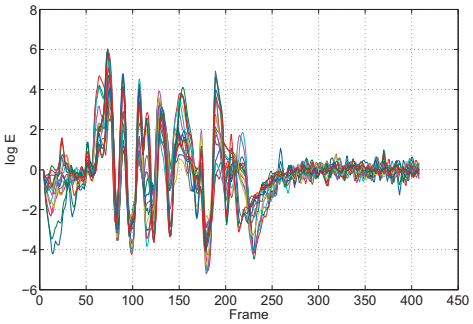

Рис. 4. Пример траекторий логарифмов энергий в критических полосах после RASTA-фильтрации

Как правило, для мужчин характерны большие значения периода основного тона (6-20 мс) по сравнению с женщинами (3-8 мс)[1]. Однако, эти диапазоны все же пересекаются, так что в некоторых случаях женскому голосу может соответствовать больший

[1]Приведенные диапазоны значений условны и могут в частных случаях сильно изменяться.

период ОТ. Поэтому наиболее сложными представляются ситуации, когда необходимо идентифицировать женщину с низким голосом или, наоборот, мужчину с высоким.

На первом рисунке представлен случай англоязычного диктора-мужчины с высокой частотой основного тона. С другой стороны, на втором рисунке мы видим изменение частоты основного тона для женщины-японки. Ее частота основного тона в среднем ниже, чем у вышепоказаного диктора-мужчины. В таких случаях правильная идентификация должна обеспечиваться за счет использования параметров, отражающих различия в структуре голосового тракта мужчин и женщин.

1.2 RASTA-PLP коэффициенты

Исходя из вышесказанного, мы включили в вектор признаков 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) - линейное предсказание с учетом особенностей слухового восприятия и RASTA ("RelAtive SpecTrA") -обработки, предназначенной для удаления из сигнала спектральных компонент, скорость изменения которых отлична от скорости изменения соответствующих компонент речи [1]. Основные этапы этой обработки перечислены ниже.

- **Разбивка на фреймы.** При работе с частотой дискретизации $f_s = 8000$ Гц использовались фреймы длиной 25 мс (200 дискретных отсчетов) с перекрытием в 10 мс (120 дискретных отсчетов).

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

- **PLP-анализ.** Частотный диапазон $[0, f_s/2]$ разбивается на 17 критических полос (critical bands). Эти полосы соответствуют равномерному разбиению частотного диапазона в bark-шкале, получаемой из линейной herz-шкалы по формуле.

$$z = 6\log(f/600 + \sqrt{(f/600)^2 + 1}\,). \tag{1}$$

  Затем подсчитывается логарифмы энергий $\log E_i$ во всех сигнальных критических полосах. Традиционно последующими этапами PLP-анализа являются умножение на кривую равной громкости и имитация закона слухового восприятия, однако в RASTA-алгоритме они выполняются после межфреймового сглаживания величин $\log E_i$, описываемого ниже.

- **RASTA-фильтрация.** Дискретная передаточная функция RASTA-фильтра имеет вид [1]

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

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

- **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 он умножается на кривую равной громкости [2], которая определяется соотношением

$$H(f) = \frac{f^4}{(f^2 + 1.6 \times 10^5)^2} \times \frac{f^2 + 1.44 \times 10^6}{f^2 + 9.61 \times 10^6}, \tag{3}$$

  где f - частота в линейном масштабе.

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

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

- **Вычисление коэффициентов линейного предсказания (КЛП).** Для вычисления КЛП порядка 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 [6].

- **Преобразование в кепстральные коэффициенты.** Кепстральные коэффициенты (т.е. обратное преобразование Фурье от АР спектра) вычисляются через рекуррентные соотношения [6]:

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

  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 $c'_1, c'_2, \ldots, c'_{10}$.

2 МОДЕЛЬ ГАУССОВЫХ СМЕСЕЙ (GMM)

Основная идея аппарата GMM состоит, как известно [5], в представлении плотности распределения вектора акустических параметров $\mathbf{x}$ в виде взвешенной суммы гауссовских плотностей распределения:

$$p(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m b(\mathbf{x}/\mu_m, \mathbf{D}_m), \tag{4}$$

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

$$b(\mathbf{x}/\mu, \mathbf{D}) = \frac{1}{\sqrt{2\pi \det D}} \exp(-0.5(\mathbf{x}-\mu)^T \mathbf{D}^{-1}(\mathbf{x}-\mu)), \quad (5)$$

Фактически представление плотности $p(\mathbf{x})$ 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.

2.1 Обучение гауссовых смесей

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 $\alpha_i, \mu_\mathbf{i}, \mathbf{D_i}, i = 1, \ldots, M$. Исходными данными для обучения является набор векторов акустических признаков $\mathbf{X} = \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_T}$.

Обучение GMM традиционно осуществляется по алгоритму EM (expectation-maximization). Существуют 2 варианта для вычисления ковариационных матриц $\mathbf{D_i}$, предполагающие их "полную" или диагональную структуру. Соответствующие итеративные соотношения приведены в работах [3] и [5] соответственно.

2.2 Проверка гипотез

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

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

$$L^{(fem.)} = \frac{1}{N} \log p(\mathbf{X}/\alpha^{(fem.)}, \mu^{(fem.)}, \mathbf{D}^{(fem.)}) = \frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{x}_i/\alpha^{(fem.)}, \mu^{(fem.)}, \mathbf{D}^{(fem.)}),$$

где $\log p(\mathbf{x}_i/\alpha^{(male)}, \mu^{(male)}, \mathbf{D}^{(male)})$ и $\log p(\mathbf{x}_i/\alpha^{(fem.)}, \mu^{(fem.)}, \mathbf{D}^{(fem.)})$ записываются в соответствии с (4).

Если $L^{(male)} > L^{(fem.)}$, выносится решение о преобладании мужского голоса. В противном случае, выносится решение о преобладании женского голоса.

3 ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Для обучения мужских и женских GMM нами были использованы доступные фрагменты баз "CSLU Multi-language" и "CSLU 22 languages". Общая продолжительность записей составила около 35 минут. В формировании записей участвовали 21

мужчина и 13 женщин. Среди языков были представлены Португальский, Английский, Немецкий, Хинди, Венгерский, Японский, Русский, Испанский. Вычисление акустических признаков осуществлялось только на вокализованных фреймах. Количество компонент гауссовых смесей было взято равным восьми. Экспериментальная проверка метода классификации была проведена на независимой русскоязычной речевой базе, записанной с участием шести мужчин и трех женщин (в общей сложности 1000 речевых файлов продолжительностью около 4 секунд каждый). Результаты проверки приведены в табл. 1. Как видно из таблицы, использование 8-компонентных гауссовых смесей с полными ковариационными матрицами обеспечивает 100%-ное определение на независимой базе. Этот результат служит основанием для продолжения тестирования алгоритма на более широких базах тестовых речевых сигналов.

ВЫВОДЫ

В данной работе предложен автоматический классификатор пола диктора на основе аппарата гауссовых смесей и RASTA-PLP обработки. Предварительное тестирование показало высокую вероятность правильного определения пола (от 98 до 100% для различных модификаций метода). В развитие данной работы необходимо перейти к тестированию алгоритма на более широких базах дикторов.

Табл. 1. Процент ошибок при тестировании на независимой базе для диагональных и полных ковариационных матриц (КМ) размерностью 8×8

	Диаг. КМ	Полн. КМ
Мужчины	2%	0%
Женщины	0%	0%

ЛИТЕРАТУРА

1. *Hermansky H., Morgan N.* RASTA processing of speech // IEEE Trans. Speech and Audio Processing.– 1994.– **2**, N 6.– P. 578–589.

2. *Hermansky H.* Perceptual Linear Prediction (PLP) analysis of speech // J. Acoust. Soc. America.– 1990.– **87**.– P. 1738–1753.

3. *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.– Dalian, 2006.– P. 3376–3379.

4. *Вовк И. В., Семенов В.Ю.* Автоматическое обнаружение и распознавание сухих хрипов на основе анализа их автокорреляционной функции // Акуст. вісн..– 2005.– **8**, N 3.– С. 17–23.

5. *Reynolds D.A., Rose R.C.* Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models // IEEE Trans. Speech Audio Proces..– 1995.– **3**.– P. 72–83.

6. *Рабинер Л., Шафер Р.* Цифровая обработка речевых сигналов.– М.: Радио и связь, 1981.– 496 с.