ISSN 1028-7507 Акустичний вісник. 2009. Том 12, N 2. С. 31–38

УДК 534.78+621.391

МЕТОД ИДЕНТИФИКАЦИИ ПОЛА ДИКТОРА НА ОСНОВЕ МОДЕЛИРОВАНИЯ АКУСТИЧЕСКИХ ПАРАМЕТРОВ ГОЛОСА ГАУССОВЫМИ СМЕСЯМИ

А. Я. КАЛЮЖНЫЙ, В. Ю. СЕМЕНОВ

ГНПП "Дельта", Киев

Получено 15.09.2009

В статье предложен метод автоматической классификации речевых фрагментов по признаку "мужчина/женщина" и описаны основные этапы его алгоритмической реализации. Метод основан на моделировании плотности распределения вектора акустических признаков голоса взвешенной суммой нескольких гауссовских распределений (метод гауссовых смесей). Каждый из членов GMM соответствует некоторому подклассу множества акустических параметров голосового сигнала. В качестве вектора акустических признаков была выбрана совокупность кепстральных RASTA-PLP коэффициентов, дополненных периодом основного тона. Обучение гауссовых смесей для мужских и женских голосов проводилось по методу expectation-maximization с инициализацией согласно алгоритму K-средних. Исследована зависимость процента ошибок классификации от типа ковариационных матриц GMM и их порядков. В различных экспериментах предложенный метод показал достаточно малую вероятность ошибки классификации (от 9 до 0 %). Сделан вывод о вторичности порядка и типа GMM по сравнению с необходимостью разнообразного представления дикторов в обучающей базе речевых сигналов.

В статті запропоновано метод автоматичної класифікації мовних фрагментів за ознакою "чоловік/жінка" та описані основні етапи його алгоритмічної реалізації. Метод заснований на моделюванні щільності розподілу вектора акустичних ознак голосу зваженою сумою декількох гаусівських розподілів (метод гаусових сумішей, GMM). При цьому кожний член GMM відповідає деякому підкласу множини акустичних параметрів голосового сигналу. За вектор акустичних ознак було обрано сукупність кепстральних RASTA-PLP коефіцієнтів, доповнених періодом основного тону. Навчання гаусових сумішей для чоловічих та жіночих голосів виконувалось за методом expectation-maximization з ініціалізацією згідно алгоритму K-середніх. Досліджено залежність процента помилок класифікації від типу коваріаційних матриць GMM та їхніх порядків. У різних експериментах запропонований метод показав достатньо малу ймовірність помилки класифікації (від 9 до 0 %). Зроблено висновок щодо другорядності порядку та типу GMM у порівнянні з необхідністю різноманітного представлення дикторів у навчальній базі мовних сигналів.

The method for automatic speaker's gender classification has been proposed and its basic algorithmic stages have been described. The method is based on modeling of voice acoustic parameters distribution by a weighted sum of several Gaussian distributions (Gaussian mixture modeling, GMM). In doing so, every component of the GMM corresponds to a certain subset of voice acoustic parameters. The set of cepstral RASTA-PLP coefficients extended by the period of the basic tone has been selected as the vector of acoustic features. The male and female GMMs were trained by the expectation-maximization method initialized according to the K-means algorithm. The dependence of classification errors on the GMM types and their orders has been investigated. In different experiments, the proposed method has shown low probability of classification errors (from 9 to 0 %). This fact allows the conclusion about minor importance of the GMM order and type in comparison with a necessity of the diverse presenting of the speakers in the training data set.

ВВЕДЕНИЕ

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 при обеспечении правоохранительной деятельности, сборе информации для рекламных целей и т. п.

Упрощенная структура системы распознавания представлена на рис. 1. Как известно, ключевыми вопросами для построения любой системы распознавания являются:

1) выбор признаков, т. е. параметров, характеризующих распознаваемые объекты (в данном случае – мужские/женские голоса);
2) выбор модели, в соответствии с которой производится обучение системы распознавания и последующая классификация признаков.

Согласно схеме, на предварительном этапе из базы тестовых сигналов выделяются векторы признаков, используемые для обучения классифицирующей модели. В результате этого формируются некоторые классы или эталонные значения признаков. В процессе реальной работы проверяемый сигнал подвергается предварительной обработке (масштабированию, удалению шумов). Сравнение извлеченных из него признаков с полученными на предварительном этапе эталонными значениями в соответствии с некоторым решающим правилом дает результат классификации.

В роли вектора признаков обычно выступают кепстральные параметры, вычисляемые на каждом фрейме речевого сигнала. В задачах ра-

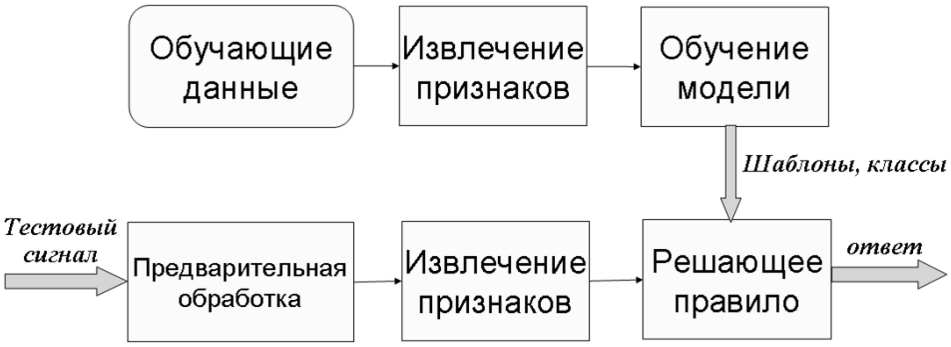

Рис. 1. Упрощенная структура системы распознавания

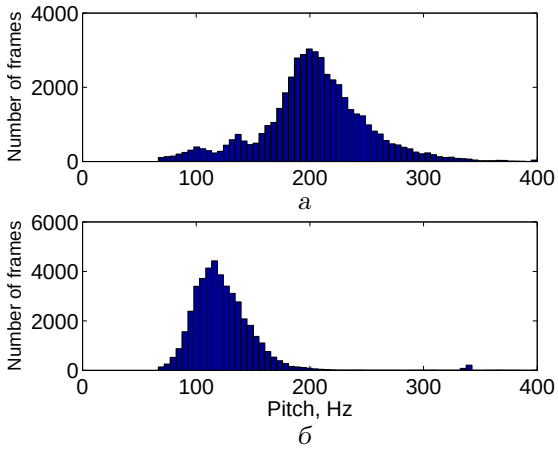

Рис. 2. Гистограммы распределения частоты основного тона:
а – для женщин; б – для мужчин

спознавания используются различные подходы к выбору классификации: гауссовы смеси (Gaussian Mixture Models – GMM), скрытые марковские модели (HMM) и др. Выбор между методами GMM и HMM зависит от того, является ли текстозависимой поставленная задача идентификации. Поскольку нас интересует не динамическая смена признаков, а интегральное преобладание одних признаков над другими, целесообразно считать, что обсуждаемая задача текстонезависима, и применять аппарат GMM.

1. ВЫЧИСЛЕНИЕ ПРИЗНАКОВ

1.1. Период основного тона

Важный признак, используемый для различения мужских и женских голосов, – период T_0 или частота $f_0 = 1/T_0$ основного тона. Этот параметр характеризует частоту колебания голосовых связок при произнесении звонких звуков. Для вычисления периода основного тона мы использовали автокорреляционный метод, описанный в работе [4].

Как правило, для мужчин характерны более низкие частоты основного тона, чем для женщин. Однако, как видно из рис. 2, эти диапазоны для различных полов пересекаются, так что в некоторых случаях женскому голосу может соответствовать меньшая частота основного тона. Поэтому наиболее сложными представляются ситуации, когда необходимо идентифицировать женщину с низким голосом или, наоборот, – мужчину с высоким. В таких случаях правильная идентификация должна обеспечиваться за счет использования параметров, отражающих различия в структуре голосовых трактов мужчин и женщин.

1.2. RASTA-PLP коэффициенты

Исходя из сказанного, по аналогии с работой [3] мы включили в вектор признаков из 10 RASTA-PLP коэффициентов, определяющих форму голосового тракта при произнесении звуков, дополненных периодом основного тона. При этом был исключен кепстральный коэффициент, отвечающий за уровень сигнала, т. е. общая размерность вектора признаков составляла 11.

Поясним, что методика анализа речевых сигналов RASTA-PLP состоит из двух частей: PLP (Perceptual Linear Prediction [2]) – линейного предсказания с учетом особенностей слухового восприятия и RASTA-обработки (от “RelAtive SpecTrA” – относительные спектры), предназначенной для удаления из сигнала спектральных компонент, скорость изменения которых отлична от скорости изменения соответствующих компонент речи [1]. Перечислим основные этапы RASTA-PLP обработки.

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).

Вычисление спектра

На каждом фрейме вычислялся квадрат модуля преобразования Фурье. Для этого речевой фрейм длиной L отсчетов предварительно дополнялся нулями до длины $L_{\text{fft}} = 2^{[(\log_2 L)]+1}$, после чего применялось окно Хемминга.

PLP-анализ

Частотный диапазон $[0, f_s/2]$ разбивался на 17 критических полос, соответствующих равномерному разбиению частотного диапазона в bark-шкале [2], получаемой из линейной Гц-шкалы по формуле

$$z = 6 \log \left[\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1} \right]. \qquad (1)$$

Каждой из полос соответствует фильтр с трапецеидальной частотной характеристикой в логарифмическом масштабе (они приведены на рис. 3).

Пусть z_0 – центр некоторой критической полосы, выраженный в барках. Тогда ее амплитудно-частотная характеристика $h(z)$ вычисляется по формуле

$$h(z) = \begin{cases} 10^{z-z_0+1/2}, & z < z_0 - \dfrac{1}{2}; \\ 1, & z_0 - \dfrac{1}{2} \le z \le z_0 + \dfrac{1}{2}; \\ 10^{-2.5(z-z_0-1/2)}, & z > z_0 - \dfrac{1}{2}. \end{cases} \qquad (2)$$

Для каждого сигнального фрейма подсчитаем суммарную энергию во всех сигнальных критических полосах:

$$\log E_i = \log \sum_{j=1}^{L_{\text{fft}}/2+1} h_i^j X_j, \qquad i = 1, \ldots, 17, \qquad (3)$$

где X – спектр мощности сигнального фрейма, полученный на предыдущем этапе; i – номер критической полосы; j – номер спектрального отсчета.

Подсчитаем логарифмы энергий $\log E_i$ во всех сигнальных критических полосах. Традиционно

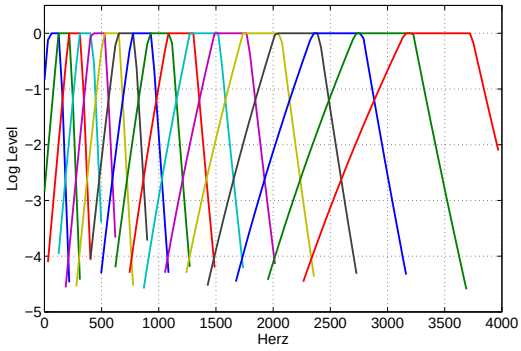

Рис. 3. Амплитудно-частотные характеристики PLP-фильтров в логарифмическом масштабе

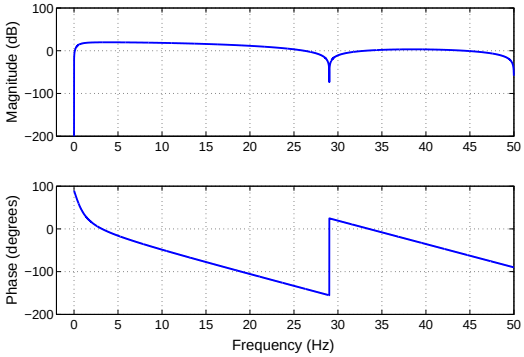

Рис. 4. Амплитудная и фазовая частотная характеристика RASTA-фильтра

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 $\log E_i$, описанного ниже.

RASTA-фильтрация

Дискретная передаточная функция RASTA-фильтра имеет вид [1]

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

Амплитудная и фазовая частотные характеристики этого фильтра представлены на рис. 4. Проанализируем его частотные свойства, предполагая, что частота обновления фреймов составляет $S_r = 100$ Гц. Корни числителя передаточной функции (4) равны $1, -1, -0.25 \pm 0.97$. Это говорит о том, что амплитудно-частотная характеристика имеет провалы при 0 $S_r/2 = 50$ и

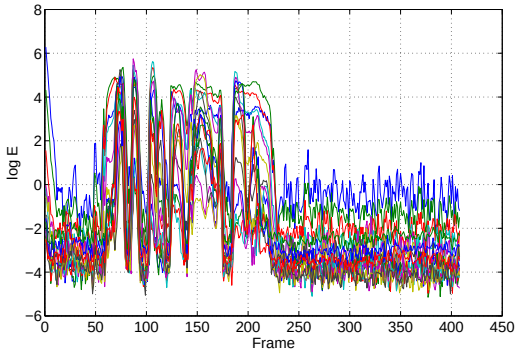

Рис. 5. Пример траекторий логарифмов энергий в критических полосах до RASTA-фильтрации

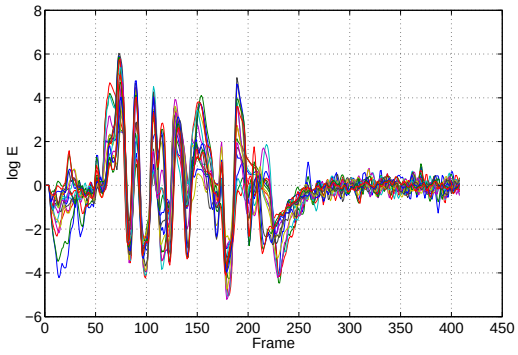

Рис. 6. Пример траекторий логарифмов энергий в критических полосах после RASTA-фильтрации

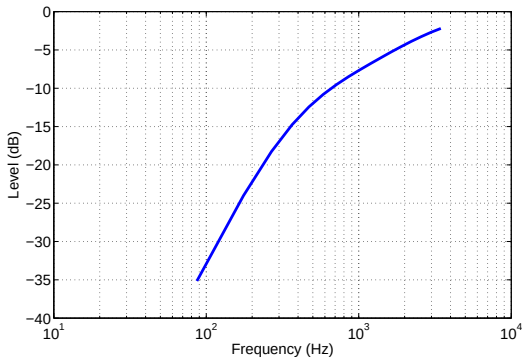

Рис. 7. Кривая равной громкости

$[\pi-\arctan(0.97/0.25)]S_r/(2\pi)=28.9$ Гц. Что касается знаменателя (4), то его полюс 0.94 соответствует константе экспоненциального накопления $0.94/(1-0.94)=16$, т. е. эффективному накоплению информации на протяжении 160 мс.

Через фильтр $R(z)$ пропускается каждая из 17 спектральных траекторий $\log E_i$, полученных на предыдущем этапе. На рис. 5 и 6 показаны траектории логарифмов энергий в критических полосах до и после RASTA-фильтрации, убирающей 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], рис. 7, которая определяется соотношением

$$H(f)=\frac{f^4}{(f^2+1.6\cdot 10^5)^2}\cdot\frac{f^2+1.44\cdot 10^6}{f^2+9.61\cdot 10^6}. \qquad (5)$$

Здесь 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 $R(k)$, $k=0,\ldots,L_{\text{fft}}-1$.

Вычисление коэффициентов линейного предсказания

Для вычисления коэффициентов линейного предсказания порядка p (в нашем случае $p=10$) необходимы первые $(p+1)$ значений автокорреляционной функции: $R(0)$, $R(1)$,..., $R(p)$. Их можно найти с помощью рекурсии Левинсона – Дарбина [10], строящейся по следующим правилам:

$$E^{(0)}=R(0);$$

$$k_i=\frac{1}{E^{(i-1)}}\left[R(i)-\sum_{j=1}^{i-1}\alpha_j^{(i-1)}R(i-j)\right],\quad 1\le i\le p;$$

$$\alpha_i^{(i)}=k_i;$$

$$\alpha_j^{(i)}=\alpha_j^{(i-1)}-k_i\alpha_{i-j}^{(i-1)},\qquad 1\le j\le i-1;$$

$$E^{(i)}=(1-k_i^2)E^{(i-1)}.$$

Окончательно коэффициенты линейного предсказания a_1, a_2,..., a_p вычисляются как

$$a_j=-\alpha_j^{(p)},\qquad j=1,\ldots,p.$$

Преобразование в кепстральные коэффициенты

Кепстральные коэффициенты (т. е. обратное преобразование Фурье от логарифма спектра сигнала) вычисляются через рекуррентные соотношения [10]:

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

На завершающем этапе коэффициенты проходят процедуру "лифтинга":

$$c'_n = n^{0.6} c_n, \qquad n = 1, \ldots, p.$$

Таким образом, итоговый вектор признаков состоит из периода основного тона T_0 и RASTA-PLP коэффициентов c'_1, c'_2,..., c'_{10}.

2. МОДЕЛЬ ГАУССОВЫХ СМЕСЕЙ (GMM)

Основная идея аппарата GMM состоит в представлении плотности распределения вектора акустических параметров $\boldsymbol{x}$ (размерностью d) в виде взвешенной суммы гауссовских плотностей распределения [6]:

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} \alpha_m b(\boldsymbol{x}/\boldsymbol{\mu}_m, \boldsymbol{D}_m), \tag{6}$$

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

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

Фактически представление плотности $p(\boldsymbol{x})$ в виде суммы M гауссианов соответствует разбиению множества акустических параметров на M подклассов [6]. Такой подход схож с идеей векторного квантования, однако более гибок.

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

Обучение гауссовых смесей

GMM должны быть независимо обучены для каждого из альтернативных классов дикторов – мужского и женского. Это означает, что для него должен быть найден свой набор параметров α_i, $\boldsymbol{\mu}_i$, $\boldsymbol{D}_i$, $i=1, \ldots, M$. Исходными данными для обучения является набор векторов акустических признаков $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$.

Обучение GMM традиционно осуществляется по алгоритму EM (expectation-maximization – дословно: максимизации ожидания) [9]. Существуют два варианта для вычисления ковариационных матриц $\boldsymbol{D}_i$, предполагающие их "полную" или диагональную структуру. Соответствующие итеративные соотношения даны в работах [3, 6].

Приведем уравнения для итеративного вычисления параметров α_i, $\boldsymbol{\mu}_i$, $\boldsymbol{\sigma}_i$, $i=1, \ldots, M$ в случае диагональных ковариационных матриц [6]:

- обновление апостериорных вероятностей попадания в m-й класс:

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

где

$$b_m(\boldsymbol{x}_i) = \frac{\exp\left\{ -\frac{1}{2} \sum_{k=1}^{d} \frac{(\boldsymbol{x}_i^k - \boldsymbol{\mu}_m^k)^2}{(\boldsymbol{\sigma}_m^k)^2} \right\}}{\prod_{k=1}^{d} \boldsymbol{\sigma}_m^k};$$

- обновление весов:

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

- обновление средних значений:

$$\boldsymbol{\mu}_m = \frac{\sum_{i=1}^{N} p(m/\boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \boldsymbol{x}_i}{\sum_{i=1}^{N} p(m/\boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma})};$$

- обновление дисперсий:

$$(\boldsymbol{\sigma}_m)^2 = \frac{\sum_{i=1}^{N} p(m/\boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}) (\boldsymbol{x}_i)^2}{\sum_{i=1}^{N} p(m/\boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma})} - (\boldsymbol{\mu}_m)^2.$$

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

Инициализация алгоритма EM

Как правило, для метода EM остро стоит проблема начальной инициализации и в специальной литературе этому вопросу уделялось большое внимание. Тем не менее, в работе [6] сделано замечание, что в конечном итоге результаты идентификации диктора не сильно зависят от способа инициализации алгоритма в процессе тренинга гауссовых смесей.

Для инициализации обучения GMM мы использовали алгоритм K-средних [8], применение которого к набору векторов акустических признаков $\boldsymbol{X}=\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T$ позволяет найти M квантов, служащих инициализацией для математических ожиданий $\boldsymbol{\mu}_m$,$m=1,\ldots,M$. Далее, отбирая вектора $\boldsymbol{x}_i$, попавшие в m-ю ячейку $\mathsf{K}^{(m)}$, получаем приближение для дисперсий:

$$(\sigma_m^k)^2=\frac{\sum\limits_{i\in\mathsf{K}^{(m)}}(\boldsymbol{x}_i^k-\boldsymbol{\mu}_m^k)^2}{N^{(m)}},\qquad k=1,\ldots,d,$$

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

Значения $\boldsymbol{\alpha}$ инициализируются как

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

Проверка гипотез

В процессе реальной работы, когда имеется набор из N наблюдений $\boldsymbol{X}=\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N$, проверка гипотез сводится к простому сравнению плотностей вероятностей, соответствующих наличию голосов дикторов каждого из полов (мужского – индекс "(mal.)" и женского – индекс "(fem.)"): $p(\boldsymbol{X}/\alpha^{(\mathrm{mal.})},\boldsymbol{\mu}^{(\mathrm{mal.})},\boldsymbol{D}^{(\mathrm{mal.})})$ и $p(\boldsymbol{X}/\alpha^{(\mathrm{fem.})},\boldsymbol{\mu}^{(\mathrm{fem.})},\boldsymbol{D}^{(\mathrm{fem.})})$. Предполагая независимость векторов наблюдений, эти величины запишем в нормированном логарифмическом масштабе:

$$L^{(\mathrm{mal.})}=\frac{1}{N}\log p\left(\boldsymbol{X}/\alpha^{(\mathrm{mal.})},\boldsymbol{\mu}^{(\mathrm{mal.})},\boldsymbol{D}^{(\mathrm{mal.})}\right)=$$

$$=\frac{1}{N}\sum_{i=1}^{N}\log p\left(\boldsymbol{x}_i/\alpha^{(\mathrm{mal.})},\boldsymbol{\mu}^{(\mathrm{mal.})},\boldsymbol{D}^{(\mathrm{mal.})}\right),$$

$$L^{(\mathrm{fem.})}=\frac{1}{N}\log p\left(\boldsymbol{X}/\alpha^{(\mathrm{fem.})},\boldsymbol{\mu}^{(\mathrm{fem.})},\boldsymbol{D}^{(\mathrm{fem.})}\right)=$$

$$=\frac{1}{N}\sum_{i=1}^{N}\log p\left(\boldsymbol{x}_i/\alpha^{(\mathrm{fem.})},\boldsymbol{\mu}^{(\mathrm{fem.})},\boldsymbol{D}^{(\mathrm{fem.})}\right),$$

где оба логарифма записываются в соответствии с выражением (6).

Если $L^{(\mathrm{mal.})}>L^{(\mathrm{fem.})}$, выносится решение о преобладании мужского голоса. В противном случае считается, что преобладает женский голос.

3. ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Для проверки метода нами были сформированы две базы речевых сигналов, описываемые ниже.

База 1. В формировании записей участвовали 16 мужчин и 11 женщин. Среди языков были представлены русский и английский (США). Для каждого диктора бралось по 10 файлов, общая продолжительность которых составила около 8 минут для мужчин и 6 – 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 продолжительность составила по 20 минут для мужчин и для женщин (103 и 154 файла соответственно).

В первом эксперименте база 1 была взята в качестве обучающей, а база 2 – в качестве проверочной. Во втором эксперименте мы поступили наоборот: база 2 выступала как обучающая, а база 1 – как проверочная. При этом количество компонент гауссовых смесей взято равным 1, 2, 4, 8, 12, или 16.

В табл. 1 и 2 приведены проценты ошибок классификации при различных порядках гауссовых смесей, а также типов ковариационных матриц для первого и второго экспериментов соответственно. Меньшее количество ошибок при использовании базы 2 в качестве обучающей объясняется ее большим объемом и большим разнообразием дикторов по сравнению с базой 1. Кроме того, увеличение количества компонент гауссовых смесей не приводит к уменьшению ошибки распознавания. Более того, в первом эксперименте самый низкий средний процент ошибок получен при использовании лишь одного гауссиана (т. е. при моделировании пространства признаков пола одним акустическим классом). Этот результат выглядит неожиданно. Возможно, он связан с относительно небольшим объемом обучающей базы 1, что порождает необходимость дальнейшего тестирования алгоритма на более широких базах сигналов. Тем не менее, обнаруженный парадокс частично подтверждается при использовании базы 2 в качестве обучающей, а также выводами исследования [3], где модификация с двумя гауссианами обеспечивала практически такой же процент

Табл. 1. Процент ошибок для диагональных и полных ковариационных матриц различных размерностей (первый эксперимент)

Классификация	Диаг. 1	Диаг. 2	Диаг. 4	Диаг. 8	Диаг. 12	Диаг. 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 %
Классификация	Полн. 1	Полн. 2	Полн. 4	Полн. 8	Полн. 12	Полн. 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 %

Табл. 2. Процент ошибок для диагональных и полных ковариационных матриц различных размерностей (второй эксперимент)

Классификация	Диаг. 1	Диаг. 2	Диаг. 4	Диаг. 8	Диаг. 12	Диаг. 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 %
Классификация	Полн. 1	Полн. 2	Полн. 4	Полн. 8	Полн. 12	Полн. 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 %

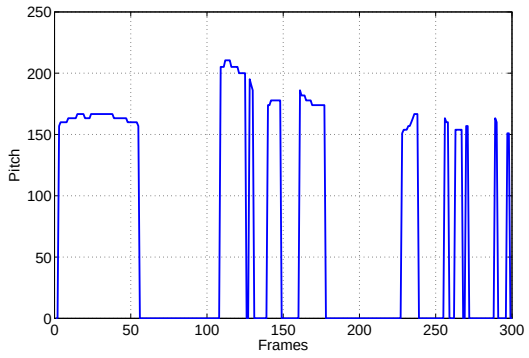

Рис. 8. Пример изменения частоты основного тона для англоязычного диктора-мужчины

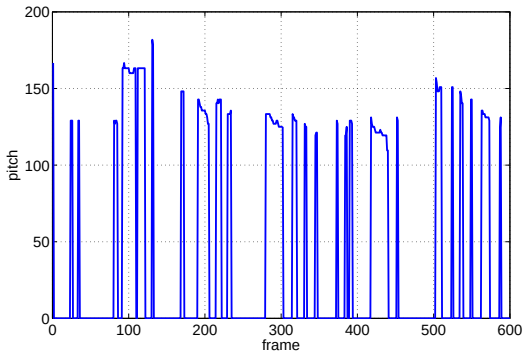

Рис. 9. Пример изменения частоты основного тона для японоязычного диктора-женщины

ошибок, как и для $M = 4, 6, 8, 10, 12, 16$.

Также заметим, что практически все ошибки в первом эксперименте наблюдались для двух дикторов – женщины-японки со средней частотой основного тона около 135 Гц и мужчины-англичанина со средней частотой около 200 Гц, которые несколько нетипичны для соответствующих полов (см. рис. 1). Примеры траекторий частоты основного тона для этих двух дикторов приведены на рис. 8 и 9.

Подводя итоги, можно заключить, что порядок и тип GMM не оказывают существенного влияния на процент ошибок при классификации голосов по половому признаку. Главным же фактором является разнообразие дикторов в обучающей базе речевых сигналов.

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 характеризуе-

тся существенно меньшими вычислительными затратами, чем при использовании полных ковариационных матриц. Заметим, что в работе [3] также был сделан выбор в пользу диагональной ковариационной матрицы размерности 8×8.

ВЫВОДЫ

1. 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.

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

3. Для правильной классификации пола диктора порядок и тип GMM оказались вторичным фактором по сравнению с необходимостью разнообразить представление голосов дикторов в обучающей базе речевых сигналов.

4. Наиболее практичными нам видятся модификации с диагональными ковариационными матрицами малого размера (например, 4×4 или 8×8), поскольку они дают приемлемый процент распознавания и характеризуется существенно меньшими вычислительными затратами, чем при использовании полных ковариационных матриц.

1. *Hermansky H., Morgan N.* RASTA processing of speech // IEEE Trans. Speech Audio Proces.– 1994.– **2**.– P. 578–589.
2. *Hermansky H.* Perceptual Linear Prediction (PLP) analysis of speech // J. Acoust. Soc. Amer.– 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 // Proc. Fifth Int. Conf. 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., Quatieri T. F., Dunn R. B.* Speaker verification using adapted Gaussian mixture models // Digit. Signal Proces.– 2000.– **10**.– P. 19–41.
6. *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.
7. *Reynolds D. A.* Experimental evaluation of features for robust speaker identification // IEEE Trans. Speech Audio Proces..– 1994.– **2**.– P. 639–643.
8. *Linde Y., Buzo A.,Gray R. M.* An algorithm for vector quantizer design // IEEE Trans. Com.– 1980.– **28**, N 1.– P. 84–95.
9. *Dempster A., Lair N., Rubin D.* Maximum likelihood from incomplete data via the EM algorithm // J. Roy. Statistic. Soc.– 1977.– **39**.– P. 1–38.
10. *Рабинер Л., Шафер Р.* Цифровая обработка речевых сигналов.– М.: Радио и связь, 1981.– 496 с.