УДК 681.3

Т. С. Пеховский, А. Ю. Сизов

СРАВНЕНИЕ РАЗЛИЧНЫХ СМЕСЕЙ ГАУССОВЫХ PLDA-МОДЕЛЕЙ В ЗАДАЧЕ ТЕКСТОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ ДИКТОРА

Исследуется актуальность использования классической смеси PLDA-моделей с распределением Гаусса в качестве априорного в пространстве i-векторов для задачи верификации диктора. Исследуются условия эксперимента, в которых это использование выгодно при существующих ограничениях размеров обучающих баз. Показано, что в рамках кроссканальной задачи использование смеси двух PLDA-моделей эффективнее, чем традиционная схема с использованием одной PLDA-модели.

Ключевые слова: *i-вектор, совместный факторный анализ, смесь PLDA-моделей, распознавание диктора.*

Введение. В последнее десятилетие активно развиваются технологии текстонезависимого распознавания личностей по голосу (дикторов). В работах Рейнольдса впервые было предложено для таких задач использовать смеси гауссовых распределений (Gaussian Mixture Models, GMM) [1, 2]. В работе [2] была показана эффективность универсальной фоновой модели (Universal Background Model, UBM), также показана эффективность MAP-адаптации (Maximum A-Posteriori Probability) модели GMM-UBM при получении модели диктора.

Модель GMM-UBM обычно обучается на большой базе дикторов, с использованием критерия максимального правдоподобия и, как правило, имеет 2048 компонент. Модель диктора здесь получается путем адаптации только средних модели GMM-UBM и последующей конкатенации отдельных компонент, с формированием при этом GMM-супервектора средних — высокоразмерного вектора признаков $m(s, h)$ для h-й сессии s-го диктора.

Работы Кенни [3—5] посвящены модели совместного факторного анализа (Joint Factor Analysis, JFA) и ее различным редуцированным версиям [6—8]. JFA — это порождающая модель, используемая с целью эффективного решения проблем междикторской и межсессионной вариативности диктора в GMM-подходе. Модель JFA можно использовать (см., например, [9]) для получения оценок верификации по критерию Неймана—Пирсона. Прогресс

современных систем верификации диктора обусловлен использованием новых низкоразмерных векторов признаков, порождаемых одной из версий JFA. В этой новой модели [10] не выполняется расщепление пространства GMM-супервектора на дикторское и канальное подпространства. Процесс обучения *T*-матрицы полной изменчивости [10] аналогичен процессу обучения матрицы собственных голосов [3], за исключением того, что

— в случае матрицы собственных голосов все сессии обучающего диктора конкатенируются для последующего обучения;

— в случае *T*-матрицы все сессии обучающего диктора расцениваются как произведенные различными дикторами.

Таким образом, вектор полной изменчивости $w(s, h)$ [10] сохраняет зависимость и от канала, и от диктора и является полным низкоразмерным аналогом супервектора $m(s, h)$. Задача расщепления пространства полной изменчивости на подпространство диктора и подпространство канала реализуется, например, с помощью линейного дискриминантного анализа (Linear Discriminate Analysis, LDA). Дальнейшее развитие текстонезависимого распознавания диктора связано большей частью с использованием векторов $w(s, h)$ в качестве входных векторов-признаков — i-векторов.

Результаты последних конкурсов по оцениванию систем распознавания дикторов (Speaker Recognition Evaluation, SRE) Национального института стандартов и технологий (National institute of Standards and Technologies, NIST) [11] показали высокую эффективность различных методов, использующих низкоразмерные i-векторы. Среди них самыми перспективными являются методы, основанные на модели вероятностного линейного дискриминантного анализа (Probabilistic LDA, PLDA) [12, 13]. В работе [12], посвященной распознаванию лиц, было представлено точное решение процедуры обучения гауссовой PLDA-модели (G-PLDA) с использованием критерия максимального правдоподобия. В работе [13] Кенни реализовал вариационное байесовское обучение PLDA-модели для верификации диктора с использованием тяжелохвостых распределений (HT-PLDA), отметив, что *t*-распределение Стьюдента должно более адекватно описывать такие негауссовы эффекты канала, как грубые искажения речи в случае записи через удаленный микрофон. Модель HT-PLDA продемонстрировала высокую эффективность при тестировании на однородном телефонном корпусе. Дальнейшее развитие подхода PLDA показало, что такую же эффективность систем верификации можно получить при использовании G-PLDA-модели, если осуществить нормализацию длины i-вектора [14].

В настоящей работе исследуются условия, при которых актуально использование классических смесей моделей G-PLDA [12], обучаемых „без учителя“ (unsupervised mixtures, U-mix) в пространстве i-векторов. U-mix позволяют осуществлять нелинейное покрытие структуры плотности данных обучающей базы, не требуя исходного знания о сегментации данных, что должно повысить эффективность системы верификации на тестовой базе, имеющей подобную структуру. По мнению авторов настоящей статьи, применение U-mix PLDA будет более актуальным в той ситуации, когда в обучающей базе априори существуют физически разнородные кластеры. Примером такой постановки задачи может являться стандартная для NIST кроссканальная задача верификации диктора, в которой обучающая база содержит данные, полученные в микрофонных и телефонных каналах.

Следует отметить, что работа [15] посвящена использованию смесей PLDA для решения кроссгендерной задачи верификации. Но, в отличие от предлагаемой нами U-mix-системы, в работе [15] обучались отдельные PLDA-системы для двух полов (компоненты смеси), обучаемые „с учителем“ (supervised mixtures, S-mix), на сегментированном материале своих полов, а смесь PLDA-моделей была реализована путем мягкого байесовского комбинирования достоверностей отдельных PLDA-систем.

В настоящей работе также ставится цель сравнить эффективность систем верификации диктора, построенных на базе моделей U-mix PLDA и на базе S-mix PLDA-моделей по схеме Кенни [16].

Обучение моделей U-mix PLDA. Поскольку в работе [12] формулы обновления гиперпараметров для G-PLDA-модели представлены без вывода, детально опишем точный вывод процедуры обучения смеси на основе критерия максимального правдоподобия.

Модель G-PLDA. Каждая из компонент рассматриваемой смеси PLDA-моделей состоит из единственной гауссовой модели фактора диктора, определенного в пространстве i-векторов. Формальное отличие от классического факторного анализа (Factor Analysis, FA) [17] заключается в том, что обучающий *s*-й диктор представлен своими *R*(*s*) сессиями, что, в свою очередь, характерно для схемы обучения PLDA-модели:

$$\begin{pmatrix} D^{(s,1)} \\ \vdots \\ D^{(s,R(s))} \end{pmatrix} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} + \begin{bmatrix} U & \cdots & 0 & V \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & U & V \end{bmatrix} \begin{pmatrix} x^{(s,1)} \\ \vdots \\ x^{(s,R(s))} \\ y^{(s)} \end{pmatrix} + \begin{pmatrix} \varepsilon^{(s,1)} \\ \vdots \\ \varepsilon^{(s,R(s))} \end{pmatrix} = \underline{D}^{(s)} = \underline{\mu} + \underline{A}\underline{z}^{(s)} + \underline{\varepsilon}^{(s)}, \qquad (1)$$

где μ — F-мерный вектор средних; $V = (F \times Q_y)$-матрица, столбцы которой можно трактовать как собственные голоса; $U = (F \times Q_x)$-матрица, ее столбцы — это собственные каналы, а шумовая (F×F)-матрица ковариации Σ — общая для всех моделей в смеси. Легко заметить, что для каждой *r*-й сессии (1) приобретает вид:

$$D^{(s,r)} = \mu + \begin{bmatrix} U & V \end{bmatrix} \begin{pmatrix} x^{(s,r)} \\ y^{(s)} \end{pmatrix} + \varepsilon^{(s,r)} = \mu + Wh^{(s,r)} + \varepsilon^{(s,r)}.$$

Здесь y, x, $\varepsilon^{(s,r)} \propto N(0,\Sigma)$ — скрытые переменные, представляющие факторы диктора, факторы канала и шум соответственно. Будем предполагать гауссов характер априорных распределений этих переменных.

Построение смеси G-PLDA моделей. Начинаем с построения функции правдоподобия смеси PLDA, состоящей из *M* моделей, используя обучающую базу из независимых дикторов, имеющих по *R*(*s*) сессий. Тогда логарифм функции правдоподобия на неполных данных есть:

$$L = \sum_{s}^{S} \ln \left\{ \sum_{m}^{M} \pi_m p_m(\underline{D}^{(s)} \mid \theta_m) \right\},$$

где π_m — веса смеси, $\theta_m = \{W_m, \mu_m, \Sigma\}$ — гиперпараметры *m*-й модели, а маргинальное правдоподобие $p_m(D^{(s)} \mid \theta_m)$ относится к отдельной вероятностной модели PLDA и выражается как

$$p_m(\underline{D}^{(s)} \mid \theta_m) = \int p_m(\underline{D}^{(s)} \mid \theta_m, \underline{z}_m) p(\underline{z}_m) d\underline{z}_m.$$

Здесь с вектором данных *s*-го диктора $\underline{D}^{(s)}$ связывается ряд бинарных скрытых переменных $\rho_m^{(s)} \in \{0,1\}$, $\sum_{m=1}^{M} \rho_m^{(s)} = 1$. Тогда параметры для этой модели смеси могут быть определены стандартным EM-алгоритмом [17] с использованием функции правдоподобия на полных данных L_c:

$$L_c = \sum_s^S \sum_m^M \rho_m^{(s)} \ln\left\{\pi_m p_m(\underline{D}^{(s)}, \underline{z}_m^{(s)} \mid \theta_m)\right\}, \tag{2}$$

где совместная вероятность:

$$p_m(\underline{D}^{(s)}, \underline{z}_m^{(s)} \mid \theta_m) = p_m(\underline{D}^{(s)} \mid \theta_m, \underline{z}_m^{(s)}) p(\underline{z}_m^{(s)}) =$$
$$= (2\pi)^{-R(s)F/2} \mid \underline{\Sigma} \mid^{-1/2} \exp\left\{-\tfrac{1}{2}(a_m^{(s)})^T \underline{\Sigma}^{-1}(a_m^{(s)})\right\} (2\pi)^{-Q/2} \exp\left\{-\tfrac{1}{2}\underline{z}_m^{(s)T} \underline{z}_m^{(s)}\right\}. \tag{3}$$

В формуле (2) $Q = Q_y + R(s)Q_x$, а $a_m^{(s)}$ есть вектор:

$$a_m^{(s)} = \left(\underline{D}^{(s)} - \underline{A}_m \underline{z}_m^{(s)} - \underline{\mu}_m\right).$$

Далее, следуя модели смеси FA [17], для математического ожидания полной функции $< L_c >$ относительно апостериорного распределения $P(z|D)$, легко получить:

$$< L_c >= \sum_s^S \sum_m^M \gamma_m(s)\Big[\ln \pi_m - \frac{1}{2} < \underline{z}_m^{(s)T} \underline{z}_m^{(s)} > - \frac{1}{2}\ln \mid \underline{\Sigma} \mid -$$
$$-\frac{1}{2}\Big\{(\underline{D}^{(s)} - \underline{\mu}_m)^T \underline{\Sigma}^{-1}(\underline{D}^{(s)} - \underline{\mu}_m) - 2(\underline{D}^{(s)} - \underline{\mu}_m)^T \underline{\Sigma}^{-1} \underline{A}_m < \underline{z}_m^{(s)} > +$$
$$+\mathrm{tr}[\underline{A}_m{}^T \underline{\Sigma}^{-1} \underline{A}_m < \underline{z}_m^{(s)} \underline{z}_m^{(s)T} >]\Big\}\Big] + \mathrm{const}.$$

Перейдем от схемы полного вектора z к представлению вектора h. Этот переход весьма облегчает последующие формулы обновления параметров в М-шаге EM-алгоритма и является очевидным, если рассмотреть скаляр под знаком экспоненты в формуле (3):

$$< (a_m^{(s)})^T \underline{\Sigma}^{-1}(a_m^{(s)}) >_{P(z|D)} = < \sum_{r=1}^{R(s)} (\xi_m^{(s,r)})^T \hat{\Sigma}^{-1}(\xi_m^{(s,r)}) >_{P(h|D)},$$

где $\xi_m^{(s,r)}$ есть вектор:

$$\xi_m^{(s,r)} = \left(D^{(s,r)} - W_m h_m^{(s,r)} - \mu_m\right).$$

Тогда математическое ожидание полной функции $< L_c >$ относительно апостериорного распределения $P(z|D)$ будет иметь вид:

$$< L_c >= \sum_s^S \sum_m^M \gamma_m(s)\Big[\ln \pi_m - \frac{R(s)}{2}\ln \mid \Sigma \mid -$$
$$-\frac{1}{2}\Big\{\sum_r^{R(s)} (D^{(s,r)} - \mu_m)^T \Sigma^{-1}(D^{(s,r)} - \mu_m) - 2\sum_r^{R(s)} (D^{(s,r)} - \mu_m)^T \Sigma^{-1} W_m < h_m^{(s,r)} > +$$
$$+\mathrm{tr}\Big[\sum_{r=1}^{R(s)} W_m{}^T \Sigma^{-1} W_m < h_m^{(s,r)} h_m^{(s,r)T} >\Big]\Big\}\Big] + \mathrm{const},$$

где компоненты парного вектора h и его ковариации должны браться из компонент полного вектора z и его ковариации [12]:

$$< h_m^{(s,r)} > \leftarrow < z_m^{(s,r)} >,$$
$$< h_m^{(s,r)} h_m^{(s,r)T} > \leftarrow < z_m^{(s,r)} z_m^{(s,r)T} >,$$

найденных, как будет описано далее, на Е-шаге EM-алгоритма. Тогда на М-шаге, в стационарной точке для функции $< L_C >$, будем иметь следующие формулы для обновления параметров:

$$\pi_m = \frac{N_m}{N} = \frac{1}{\sum_s^S \sum_m^M \gamma_m^{(s)}} \sum_s^S \gamma_m^{(s)}, \quad \mu_m = \frac{\sum_s^S \gamma_m^{(s)} \sum_r^{R(s)} (D^{(s,r)} - W_m < h_m^{(s,r)} >)}{\sum_s^S \gamma_m^{(s)} R(s)},$$

$$W_m = \left[\sum_s^S \gamma_m^{(s)} \sum_r^{R(s)} (D^{(s,r)} - \mu_m) < h_m^{(s,r)} >^T \right] \left[\sum_s^S \gamma_m^{(s)} \sum_r^{R(s)} < h_m^{(s,r)} h_m^{(s,r)T} > \right]^{-1}, \qquad (4)$$

$$\Sigma = \frac{\operatorname{diag}\left[\sum_s^S \sum_m^M \gamma_m^{(s)} \sum_r^{R(s)} < \left(D^{(s,r)} - W_m h_m^{(s,r)} - \mu_m\right)\left(D^{(s,r)} - W_m h_m^{(s,r)} - \mu_m\right)^T > \right]}{\sum_s^S \sum_m^M \gamma_m^{(s)} R(s)}.$$

Заметим, что в настоящей работе везде используется шумовая матрица ковариации Σ — общая для всех анализаторов. В формуле (4) представлен ее диагональный случай. E-шаг EM-алгоритма для смеси PLDA-моделей стандартен, так как он будет выполнен в представлении полного вектора z. На этом шаге [17] необходимо найти апостериорное распределение

$$< \underline{z}_m^{(s)} >= \underline{\Sigma}_m^{(Z)} \, \underline{A}_m{}^T \, \underline{\Sigma}^{-1} (\underline{D}(s) - \underline{\mu}_m)$$

и соответствующую матрицу:

$$< \underline{z}_m^{(s)} \, \underline{z}_m^{(s)T} >= \underline{\Sigma}_m^{(z)} + < \underline{z}_m^{(s)} >< \underline{z}_m^{(s)T} >,$$

где апостериорная матрица ковариации для обобщенного скрытого вектора z есть

$$\underline{\Sigma}_m^{(z)} = (\underline{I} + \underline{A}_m{}^T \, \underline{\Sigma}^{-1} \, \underline{A}_m)^{-1},$$

I — единичная матрица.

Также необходимо найти $\gamma_m^{(s)}$ (responsibilities) — апостериорное распределение для набора скрытых переменных $\rho_m^{(s)}$, обслуживающих смесь [17]:

$$\gamma_m^{(s)} = \frac{\rho_m^{(s)}}{\sum_k^M \rho_k^{(s)}} = \frac{\pi_m p_m(\underline{D}^{(s)})}{\sum_k^M \pi_k p_k(\underline{D}^{(s)})} = \frac{\pi_m \int p_m(\underline{D}^{(s)} \mid z) p(z) dz}{\sum_k^M \pi_k \int p_k(\underline{D}^{(s)} \mid z) p(z) dz},$$

находим точное значение маргинального правдоподобия (evidence):

$$p_m(\underline{D}^{(s)}) = \int p_m(\underline{D}^{(s)}, z) dz = \int p_m(\underline{D}^{(s)} \mid z) p(z) dz =$$

$$= (2\pi)^{-FR(s)/2} \mid \underline{C}_m \mid^{-1/2} \exp\left\{ -\tfrac{1}{2} (\underline{D}^{(s)} - \underline{\mu}_m)^T \, \underline{C}_m^{-1} (\underline{D}^{(s)} - \underline{\mu}_m) \right\} \qquad (5)$$

(здесь и далее для удобства записи будем опускать θ_m).

И, таким образом, выражение для логарифма ответственностей:

$$\ln \rho_m^{(s)} = \ln(\pi_m) - \frac{1}{2} \ln \left| \underline{C}_m \right| - \frac{1}{2} (\underline{D}^{(s)} - \underline{\mu}_m)^T \, \underline{C}_m^{-1} (\underline{D}^{(s)} - \underline{\mu}_m) + \text{const},$$

где матрица ковариации $\underline{C}_m$ в (5), после взятия интеграла для вектора диктора $\underline{D}^{(s)}$, состоящего из $R(s)$ сессий, может быть представлена как:

$$\underline{C}_m = \underline{\Sigma} + \underline{A}_m \underline{A}_m{}^T =$$

$$= \begin{bmatrix} \Sigma & & & \\ & \Sigma & & \\ & & \ddots & \\ & & & \Sigma \end{bmatrix} + \begin{bmatrix} U_m U_m{}^T + V_m V_m{}^T & V_m V_m{}^T & \cdots & V_m V_m{}^T \\ V_m V_m{}^T & U_m U_m{}^T + V_m V_m{}^T & \cdots & \vdots \\ \vdots & \vdots & \ddots & V_m V_m{}^T \\ V_m V_m{}^T & \cdots & V_m V_m{}^T & U_m U_m{}^T + V_m V_m{}^T \end{bmatrix}.$$

Обращение матриц ковариации $\underline{C}_m$ и $\underline{\Sigma}_m^{(z)}$ представляет при точном выводе определенную трудность. Но их обращение может быть сведено к обращению отдельных блоков.

Стадия верификации. *Случай U-mix PLDA*. Оценка PLDA для смеси имеет ту же структуру, что и оценка для отдельной PLDA-модели [13]:

$$\text{Score} = \ln \frac{P(D_1, D_2 \mid T)}{P(D_1 \mid I) P(D_2 \mid I)},$$

где выражение для маргинального правдоподобия в числителе (случай $R(s)$=2) и двух — в знаменателе (случай $R(s)$=1) посчитано, в отличие от [13], точно:

$$P(\underline{D}^{(s)}) = \sum_m^M \pi_m \int p_m(\underline{D}^{(s)} \mid z) p(z) dz =$$

$$= \sum_m^M \pi_m \left[(2\pi)^{-R(s)F/2} \mid \underline{C}_m \mid^{-1/2} \exp\{-\tfrac{1}{2} (\underline{D}^{(s)} - \underline{\mu}_m)^T \underline{C}_m^{-1} (\underline{D}^{(s)} - \underline{\mu}_m)\} \right]$$

и, согласно (1), представляет собой достоверность смеси PLDA-моделей.

Случай S-mix PLDA. Представим реализацию S-mix PLDA по Кенни [16], состоящую из M отдельных PLDA-моделей:

$$\text{Score} = \ln \frac{P(D_1, D_2 \mid T)}{P(D_1, D_2 \mid I)} = \ln \frac{\sum_m P(D_1, D_2 \mid m, T) P(m \mid T)}{\sum_{m,m'} P(D_1 \mid m, I) P(m \mid I) P(D_2 \mid m', I) P(m' \mid I)} =$$

$$= \ln \frac{\sum_m P(D_1, D_2 \mid m, T) P(m \mid T)}{\sum_{m,m'} Q^{(m,m')} P(D_1 \mid m, I) P(D_2 \mid m', I)},$$

где априорные распределения для целевых дикторов и „самозванцев“ (imposters) выбираются равными для каждой m-й компоненты смеси Кенни [16]:

$$P(m \mid T) = P(m \mid I) = 1 / M,$$

$$Q^{(m,m')} = P(m \mid I) P(m' \mid I) = 1 / M^2.$$

Таким образом, это можно рассматривать как вариант байесовского комбинирования отдельных PLDA-систем на стадии верификации.

Эксперименты. *Предобработка речевого сигнала.* Все записи были сегментированы на участки „речь“ и „пауза“. Участки „пауза“ затем были удалены из записей. В экспериментах использовались 39-мерные мэл-частотные кепстральные коэффициенты (mel-frequency cepstral coefficients, MFCC) [1]. MFCC-векторы состояли из 13 кепстральных коэффициентов, их первых и вторых производных, вычисляемых по 5 соседним кадрам. Использовались кадры с окном 22 мс и со сдвигом окна в 11 мс. Каждый кадр был преэмфазирован [1] и домножен на окно Хэмминга. Также везде применялась стандартная процедура вычитания кепстрального среднего из кепстральных коэффициентов.

Универсальная фоновая модель (UBM). Использовалась гендернезависимая UBM, имеющая 512 компонент и полученная с помощью EM-обучения на основе критерия максимального правдоподобия на телефонных базах NIST SRE 1998—2008 годов (все языки, оба пола). Системы PLDA обучались на записях голосов 4329 мужчин и женщин. Использовалась диагональная, а не полноковариационная GMM-UBM.

Кроссканальный экстрактор i-векторов. В кроссканальной задаче необходимо использовать универсальный экстрактор i-векторов, который бы мог адекватно работать как в телефонном, так микрофонном каналах. Здесь проблемой является несбалансированность количества записей в телефонном и микрофонном каналах. Последних в несколько раз меньше в базах NIST, чем первых. В этом случае, как предложено в работе [18], используется универсальный экстрактор i-векторов, который бы подходил как для микрофонных записей речи, так и для телефонных. Он основан на отдельных оценках максимального правдоподобия двух T-матриц полной изменчивости. Математически это можно выразить для дикторо- и каналозависимого супервектора μ следующим образом:

$$\mu = \mu_0 + T'w' + T''w''. \qquad (6)$$

В настоящей работе телефонная T' матрица с 400 базисными столбцами обучена на 11 256 телефонных записях из NIST 2002/2003/2004/2005/2006/2008 от 1250 дикторов-мужчин (только английский язык). Микрофонная T'' матрица той же размерности обучалась на 4705 микрофонных записях из NIST 2005/2006/2008 от 203 дикторов-мужчин (только английский язык), согласно [18]. Таким образом была решена проблема значительной несбалансированности наборов телефонных и микрофонных записей. После оценки T'' и T' конкатенируются, чтобы получить смешанную T-матрицу:

$$\mu = \mu_0 + Tw, \qquad (7)$$

где w-векторы есть интересующие нас итоговые i-векторы. Таким образом, используется кроссканальный экстрактор i-векторов размерности с 700 базисными столбцами.

Однородный экстрактор i-векторов. В кроссканальной задаче также будет использоваться обычный экстрактор i-векторов (6), но обученный только на телефонных записях, назовем его однородным экстрактором i-векторов. Такой необычный, на первый взгляд, выбор объясняется следующими причинами. Апостериорное распределение i-векторов обучающей базы экстрактора i-векторов (7), согласно JFA, всегда будет близко к его априорному $N(0,1)$. Таким же распределение i-векторов будет и для любой другой базы, близкой по условиям записи к обучающей (по каналу, по полу, по языку и т.д.). Но, как показали эксперименты, при существенном рассогласовании базы обучения и тестовой базы всегда наблюдается существенный сдвиг центра распределения i-векторов тестовой базы относительно нуля. Это приводит к деградации равновероятной ошибки первого и второго рода (Equal Error Rate, EER) системы, основанной на одной PLDA-модели. Но для случая обучения, например, двух PLDA моделей на двух физически явных кластерах (например, каналы в кроссканальной задаче) такое поведение однородного экстрактора будет способствовать разделению кластеров в пространстве i-векторов. Идея заключается в том, что таким образом улучшаются условия применения смеси PLDA-моделей в пространстве i-векторов, которое изначально более подходит под одну модель. Кроме того, будет использоваться однородный телефонный экстрактор i-векторов T'.

Переход в LDA-пространство. Как уже было отмечено выше, JFA-экстрактор i-векторов генерирует i-векторы, содержащие информацию как о дикторе, так и о канале. Поэтому еще одним условием, способствующим успешному применению смеси PLDA, будет переход от входных i-векторов к их проекциям, получаемым в результате LDA-преобразования. Это позволяет:

— уменьшить канальный шум;

— получить добавочную редукцию размерности входных векторов.

Такая верификационная схема $\mathrm{TV} \to \mathrm{LDA} \to \mathrm{PLDA}$ была успешно применена в различных работах по верификации диктора, а именно в кроссгендерных [15] и кроссканальных [16, 19] задачах. Метод LDA широко используется для редукции размерности в задачах классификации. В нашей работе LDA-преобразование редуцирует i-векторы до 200-мерного пространства, заполненного собственными векторами, соответствующими самым большим собственным значениям следующей обобщенной задачи о собственных значениях λ и собственных векторах x:

$$S_b x = \lambda S_w x, \tag{8}$$

где S_b и S_w — соответственно матрицы межклассовой и внутриклассовой вариативности. После решения обобщенной задачи (8) получаем LDA-матрицу, которую применяем к i-векторам в обучающих и тестовых базах. Были построены две LDA-матрицы. В случае кроссканального экстрактора обучалась LDA-матрица размерностью 700×200 на данных обучения этого экстрактора, в случае однородного экстрактора — LDA-матрица размерностью 400×200 только на 11 256 телефонных записях, использованных для обучения однородного экстрактора.

LDA-проекции i-векторов затем подвергались процедуре нормализации, согласно [14], но только для тестовой базы (U-L-G конфигурация в терминах [14]). Эта нормализация состоит в проектировании LDA-векторов на единичную сферу.

Условия обучения. Обучались две модели S-mix G-PLDA (M=2, 3) и две U-mix G-PLDA (M=1, 2). Для модели S-mix PLDA (M=3) независимо были обучены (везде — только английский язык):

— Phone-PLDA — модель, обученная на 11 256 телефонных записях из NIST 2002/2003/2004/2005/2006/2008 от 1250 дикторов-мужчин;

— Mic-PLDA — модель, обученная на 4705 микрофонных записях из NIST 2005/2006/2008 от 203 дикторов-мужчин;

— CI-PLDA — каналонезависимая PLDA-модель, обученная на совокупном наборе данных систем Phone-PLDA и Mic-PLDA.

При обучении возникает проблема сильной несбалансированности наборов телефонных и микрофонных записей NIST. Авторы решили эту проблему, взяв из 11 256 только 5000 телефонных записей дикторов, которые были представлены в микрофонном канале, и добавив к этому набору все записи по микрофонному каналу. Так же, как и в работе [16], модель S-mix PLDA (M=3) выполнена с помощью комбинирования этих трех моделей на стадии получения оценок, а S-mix PLDA (M=2) состояла из комбинации двух систем — Phone-PLDA и Mic-PLDA. Обучение компонент проводилось согласно вариационному байесовскому выводу Кенни [13]. Модели U-mix PLDA (M=1, 2) обучались на всем смешанном наборе данных двух систем Phone-PLDA и Mic-PLDA. Везде количество столбцов матрицы собственных голосов V для всех PLDA-моделей было $Q_y = 200$, а U=0. Везде в целях ускорения сходимости при обучении на основании максимального правдоподобия добавлялись итерации минимизации дивергенции Кульбака—Лейблера фазы обучения по Кенни [13]. Шумовая матрица ковариации Σ в (4) для всех случаев имела полноковариационный вид.

Результаты тестирования для кроссканала (det3). Результаты сравнения моделей U-mix и S-mix PLDA относительно результатов основного (core-core) теста на мужских голосах базы NIST SRE 2010 для кроссканальной задачи (det3) [11] представлены в табл. 1. Для оценки эффективности систем использовались ошибка EER и новый нормализованный минимум функции стоимости обнаружения NIST (Minimum Detection Cost Function, minDCF) как метрика [11].

Таблица 1

Система	*M*=1	*M*=2	*M*=3
S-mix G-PLDA Кроссканальный экстрактор	—	4,31 % [0,598]	3,83 % [0,577]
U-mix G-PLDA Кроссканальный экстрактор	3,82 % [0,579]	3,70 % [0,535]	—
U-mix G-PLDA Однородный экстрактор	4,06 % [0,601]	3,22 % [0,525]	—

Из табл. 1 следует, во-первых, что модель S-mix G-PLDA лучше всего работает при *M*=3 и осуществляет относительную редукцию EER системы на 11 % при *M*=2, а во-вторых, что модель U-mix G-PLDA при *M*=2 немного выигрывает (EER=3,70 %) у лучшей S-mix-системы при *M*=3 (EER=3,83 %) даже при использовании кроссканального экстрактора. Наконец, лучшей (EER=3,22 %) оказалась модель S-mix G-PLDA при *M*=2, использующая однородный экстрактор.

Результаты тестирования для телефонного канала (det5). Результаты сравнения систем верификации, полученных на неконтролируемой смеси PLDA-моделей, для однородного (телефон) по каналу условия (det5) представлены в табл. 2. Целью эксперимента было выяснить, можно ли наблюдать на однородном корпусе (телефон, мужчины, английский язык) структуру плотности, соответствующую выбору более чем одной модели G-PLDA. Из табл. 2 видно, что S-mix G-PLDA при *M*=2 существенно проигрывает (EER=3,97 %) системе G-PLDA (EER=3,69 %).

Таблица 2

Система	*M*=1	*M*=2
U-mix G-PLDA Однородный экстрактор	3,69 % [0,532]	3,97 % [0,585]

Обсуждение. Как ожидалось, идея однородного экстрактора оказалась весьма полезной для использования моделей U-mix PLDA. Однородный экстрактор породил на тестовой базе det3 такую же двухкластерную (телефон—микрофон) структуру плотности в пространстве i-векторов, что и в обучающем множестве. Это непосредственно следует из сравнения 2-й и 3-й строк табл. 1, видно, что в случае U-mix G-PLDA при *M*=2 во время обучения на основе максимального правдоподобия произошел захват смесью этой структуры, что положительно повлияло на эффективность этой системы (EER=3,22 %) и негативно — на эффективность системы на основе модели U-mix G-PLDA при *M*=1 (EER возрос с 3,82 до 4,06 %). Последнее свидетельствует о несоответствии структуры данных, порожденной однородным экстрактором, модели одной G-PLDA. Напротив, как следует из табл. 2, в случае однородного тестового условия (det5) эта структура, порожденная однородным экстрактором, соответствует одной модели G-PLDA. Можно сказать, что на текущий момент количество дикторов в доступных речевых базах недостаточно для эффективного использования смесей PLDA-моделей при *M*>1 в случае однородной базы данных. Таким образом, проведенные тестовые эксперименты показывают эффективность подхода моделей U-mix PLDA для кроссканальной задачи верификации диктора, которая превосходит по эффективности модель S-mix G-PLDA [16].

Заключение. В статье предложено использовать модель U-mix PLDA для решения кроссканальной задачи верификации диктора. Проведенные эксперименты на данных NIST SRE 2010 позволяют сделать следующие выводы.

1. На однородных базах данных использовать более одной модели нецелесообразно, даже в пространстве LDA-векторов, так как существующие обучающие базы на данный момент не обладают достаточным количеством дикторов.

2. На кроссканальной задаче смеси PLDA моделей можно успешно применять, но в пространстве LDA-векторов и при использовании однородного экстрактора.

3. Схема однородного экстрактора в совокупности со смесью двух моделей оказывает существенную конкуренцию схеме кроссканального экстрактора с одним гауссовым анализатором в стандартной кроссканальной задаче NIST.

В будущем планируется реализовать модель U-mix G-PLDA при использовании полной байесовской структуры. Это позволит автоматически определять релевантную размерность матриц факторов диктора и канала, а также количество компонент смеси для обучающей базы.

Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации.

СПИСОК ЛИТЕРАТУРЫ

1. *Reynolds D. A., Rose R. C.* Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Trans. Speech Audio Process. 1995. N 3. P. 72—83.

2. *Reynolds D. A., Quatieri T. F., Dunn R. B.* Speaker Verification Using Adapted Gaussian Mixture Models // Digit. Signal Process. 2000. N 10. P. 19—41.

3. *Kenny P.* Joint factor analysis of speaker and session variability: Theory and algorithms // Technical report CRIM-06/08-13. 2005.

4. *Kenny P., Boulianne G., Ouellet P., Dumouchel P.* Joint factor analysis versus eigenchannels in speaker recognition // IEEE Trans. Audio, Speech, Lang. Process. 2007. Vol. 15. P. 1435—1447.

5. *Kenny P., Ouellet P., Dehak N., Gupta V., Dumouchel P.* A Study of Inter-Speaker Variability in Speaker Verification // IEEE Trans. Audio, Speech and Lang. Process. 2008. Vol. 16. P. 980—988.

6. *Vogt R., Sridharan S.* Explicit modeling of session variability for speaker verification // Comput. Speech and Lang. 2008. Vol. 22. P. 17—38.

7. *Burget L., Matejka P., Glembek O., Cernocky J.* Analysis of feature extraction and channel compensation in GMM speaker recognition system // IEEE Trans. on Audio, Speech and Lang. Process. 2007. Vol. 15. P. 1979—1986.

8. *Pekhovsky T., Oparin I.* Eigen Channel Method for Text-Independent Russian Speaker Verification // Proc. of the XII Intern. Conf. “Speech and Comput.” SpeCom'08. Moscow, Russia, 2008. P. 385—390.

9. *Glembek O., Burget L., Brummer N., Kenny P.* Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis // IEEE Int. Conf. on Acoust., Speech, and Signal Process. Taipei, Taiwan, 2009.

10. *Dehak N., Kenny P., Dehak R., Dumouchel P., Ouellet P.* Front-end factor analysis for speaker verification // IEEE Trans. on Audio, Speech, and Lang. Process. 2010. Vol. 19. P. 788—798.

11. [Электронный ресурс]: <http://www.itl.nist.gov/iad/mig/tests/sre>.

12. *Prince S. J. D., Elder J. H.* Probabilistic linear discriminant analysis for inferences about identity // Proc. 11th Intern. Conf. on Comput. Vision. Rio de Janeiro, Brazil, 2007. P. 1—8.

13. *Kenny P.* Bayesian speaker verification with heavy tailed priors // Proc. Odyssey Speak. and Lang. Recognit. Workshop. Brno, Czech Republic, 2010.

14. *Garcia-Romero D., Espy-Wilso C. Y.* Analysis of i-vector length normalization in speaker recognition systems // Proc. of Interspeech. Florence, Italy, 2011. P. 249—252.

15. *Senoussaoui M., Kenny P., Brummer N., Villiers E., Dumouchel P.* Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition // Proc. of Interspeech. Florence, Italy, 2011. P. 25—28.

16. *Simonchik K., Pekhovsky T., Shulipa A., Afanasev A.* Supervised Mixture of PLDA Models for Cross-Channel Speaker Verification // Proc. Interspeech. Portland, USA, 2012.

17. *Tipping M., Bishop C. M.* Mixtures of probabilistic principal component analyzers // Neural Comput. 1999. Vol. 11. P. 443—482.

18. *Senoussaoui M., Kenny P., Dehak N., Dumouchel P.* An i-vector extractor suitable for speaker recognition with both microphone and telephone speech // Proc. Odyssey Speak. Recognit. Workshop. Brno, Czech Republic, 2010.

19. *Senoussaoui M., Kenny P., Dumouchel P., Castaldo F.* Well-calibrated heavy tailed Bayesian speaker verification for microphone speech // Proc. ICASSP. Prague, Czech Republic, 2011.

Сведения об авторах

Тимур Сахиевич Пеховский — канд. физ-мат. наук; ООО „ЦРТ-инновации“, Санкт-Петербург; ведущий научный сотрудник; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; доцент; E-mail: tim@speechpro.com

Александр Юрьевич Сизов — студент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: sizov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.