3. *Настасенко М. В., Дырмовский Д. В.* Эффективное использование речевой информации и биометрических технологий в силовых структурах // Вестн. МГТУ им. Н. Э. Баумана. Сер. „Приборостроение“. 2011. Вып. № 3. С. 18—25.

4. *Матвеев Ю. Н.* Технологии биометрической идентификации личности по голосу и другим модальностям // Вестн. МГТУ им. Н. Э. Баумана. Сер. „Приборостроение“. 2012. № 3 (3). С. 46—61.

5. *Дырмовский Д. В., Коваль С. Л.* Особенности человеко-машинного интерфейса современных систем биометрической идентификации // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 66—74.

Сведения об авторах

Дмитрий Викторович Дырмовский — ООО „ЦРТ“, Санкт-Петербург; директор филиала; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; соискатель; E-mail: ddv@speechpro.com

Сергей Львович Коваль — канд. техн. наук, доцент; ООО „ЦРТ“, Санкт-Петербург; главный эксперт; E-mail: koval@speechpro.com

Михаил Васильевич Хитров — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; генеральный директор; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; заведующий кафедрой; E-mail: khitrov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 004.93+57.087.1

Ю. Н. Матвеев, А. К. Шулипа

АНАЛИЗ ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ МНОГООБРАЗИЙ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ДИКТОРОВ

Исследованы особенности применения методов обучения на основе многообразий, широко используемых в приложениях по распознаванию изображений, для решения задач распознавания личностей по голосу (дикторов). Проанализированы результаты экспериментов по использованию таких методов.

Ключевые слова: *обучение на основе многообразий, распознавание диктора.*

Введение. Алгоритмы машинного обучения на основе многообразий [1] пока мало применяются в системах распознавания дикторов. Для текстонезависимого распознавания дикторов, как правило, используются методы, базирующиеся на моделировании статистических распределений речевых признаков на основе смесей гауссовых распределений, GMM [2]. Согласно оценке Национального института стандартов и технологий США (NIST), компании, занимающие лидирующие позиции в распознавании дикторов, реализуют и совершенствуют свои алгоритмы в рамках подходов на основе гауссовых смесей [3, 4]. Тем не менее в некоторых работах [5, 6] предпринимались попытки использования алгоритмов машинного обучения на основе многообразий для решения задач распознавания дикторов.

В настоящей работе рассмотрены такие алгоритмы, описываются результаты их использования и делается заключение о возможности их применения для решения задач распознавания дикторов.

Метод диффузных карт. В работе [5] решалась задача текстонезависимой идентификации дикторов. Для отображения статистических моделей речевых признаков на низкоразмер-

ное пространство использовался метод диффузных карт (Diffusion Maps). Предложенная в [5] система идентификации была реализована в три этапа.

1. *Извлечение речевых признаков на произнесениях дикторов.* В качестве речевых признаков были выбраны мел-частотные кепстральные коэффициенты (mel-frequency cepstral coefficient, MFCC) [8], для их расчета использовалась стандартная процедура [9]. Помимо кепстральных коэффициентов вычислялись их производные. Размерность вектора признаков складывалась из 13 коэффициентов и 13 производных, что соответствовало 26-мерным векторам. В итоге каждое произнесение представлялось в виде последовательности векторов.

Статистическая модель распределения признаков на произнесении строилась следующим образом: для каждой компоненты вектора признаков на всем произнесении определялись среднее, дисперсия, минимум и максимум (только для кепстральных коэффициентов). В результате каждое произнесение описывалось статистической моделью в виде 78-мерного вектора (26 средних + 26 дисперсий + 26 минимальных-максимальных значений).

2. *Отображение обучающей выборки данных в низкоразмерное пространство с использованием метода диффузных карт.* В новом пространстве элементы данных кластеризуются в соответствии с их принадлежностью дикторам. После расчета речевых признаков и статистических моделей производилось отображение 78-мерных векторов, соответствующих моделям каждого произнесения, в низкоразмерное пространство. Для преобразования пространства был применен метод диффузных карт. Причем рассматривались симметричный и несимметричный (случайный выбор) варианты расчета диффузионных матриц (подробней см. [5]).

3. *Проецирование тестового произнесения в низкоразмерное пространство* с использованием геометрических гармоник [7], затем выявление методом k ближайших соседей (k-nearest neighbor, k-NN) принадлежности произнесения к какому-либо из дикторов обучающей выборки.

На стадии классификации для тестового произнесения находилась статистическая модель, которая отображалась в низкоразмерное пространство, с этой целью использовалась формула, позволяющая выразить координаты тестового произнесения в новом пространстве на основе базиса собственных векторов, полученного на этапе обучения. Классификация тестового произнесения в низкоразмерном пространстве проводилась методом k-NN при k=10.

В работе также были исследованы результаты использования классификаторов на основе смеси гауссовых распределений (пятикомпонентная смесь) и k-NN (в этих случаях классификация проводилась без предварительной редукции пространства).

Для экспериментов была выбрана база YOHO, в которой содержатся произнесения 106 дикторов-мужчин и 32 женщин, длительность чистой речи не превышает 2—3 с [10]. Дикторы для идентификации выбирались случайным образом, поэтому значения каждого параметра, полученные в нескольких попытках при фиксированном числе дикторов, усреднялись, чтобы результат зависел только от числа дикторов в наборе.

Тестирование проводилось при двух вариантах тестовой и обучающей выборки, в первом случае объем базы тестирования был в 9 раз меньше объема базы обучения (табл. 1).

Таблица 1

Результаты идентификации (%) на наборе произнесений от различного числа дикторов

Число дикторов	Метод			
	диффузных карт		GMM	k-NN
	несимметричный	симметричный		
2	100	100	100	99,2
3	99,5	99,3	99,5	98,2
5	99,4	99,1	99,5	97,7
10	97,8	96,9	98,1	95,1
20	94,4	93,2	97,5	92,0

Во втором случае тестовая база была в 4 раза меньше базы обучения (табл. 2). Полученные результаты показывают, что использование предварительной нелинейной редукции повышает эффективность текстонезависимой идентификации в случае, когда база обучения меньше тестовой.

Таблица 2

Результаты идентификации (%) на наборе произнесений от различного числа дикторов

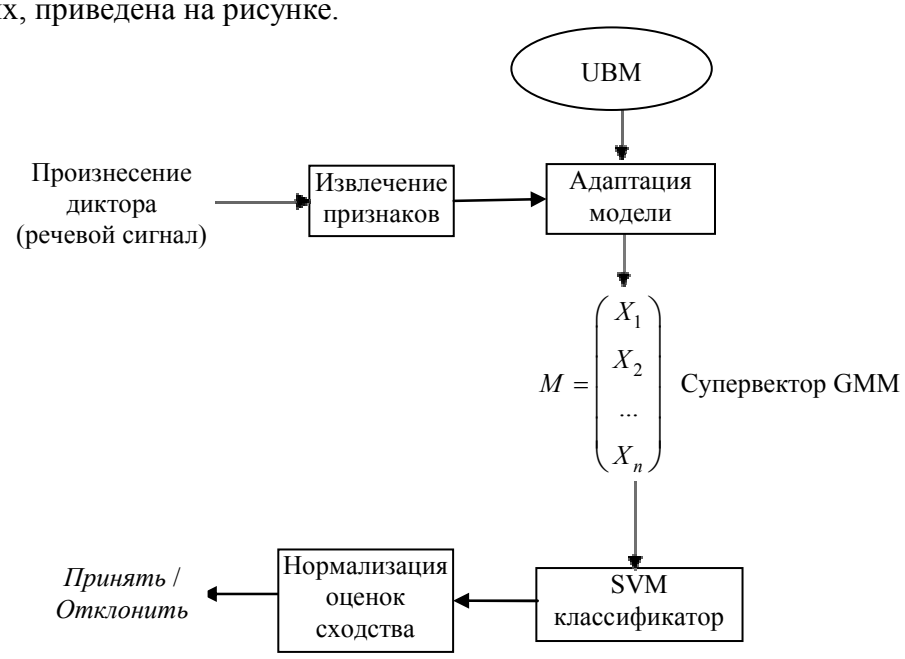
Число дикторов	Метод			
	диффузных карт		GMM	*k*-NN
	несимметричный	симметричный		
2	98,2	99,1	97,4	97,9
3	97,8	98,7	95,2	97,8
5	96,5	96,0	92,1	93,4
10	92,5	91,9	87,4	89,0
20	86,4	84,6	83,3	84,5

Уровень ошибки идентификации при редукции пространства с 78-мерного до 3-мерного примерно одинаков, что свидетельствует о применимости метода диффузных карт для выделения значимых дикторозависимых признаков.

Следует, однако, отметить, что выбранная для исследований речевая база записывалась при использовании одного и того же микрофона, поэтому вариативность, связанная с влиянием эффектов канала, сведена к минимуму, это позволило достичь сравнительно высокого качества идентификации (в среднем более 90 %).

Методы Isomap и Laplacian Eigenmaps. В работе [6] исследовалась возможность применения методов нелинейной редукции пространства к текстонезависимой верификации диктора. Топологическая структура данных моделировалась алгоритмами Isomap и Laplacian Eigenmaps, что позволило сократить размерность входного пространства данных в четыре раза без снижения качества верификации.

Структурная схема системы верификации диктора GMM-SVM, которая применялась в исследованиях, приведена на рисунке.



Это стандартная структура системы верификации диктора [2], в которой в качестве входных векторов признаков используются супервекторы GMM-UBM [11], отражающие структуру произнесений, а в качестве бинарного классификатора используется машина опорных векторов (Support Vector Machine, SVM) [2]. В системе сначала выполняется предвари-

тельная обработка тестового произнесения для выделения признаков, построение супервектора GMM, а затем выполняется классификация в модуле SVM, где принимается решение о принадлежности тестового и эталонного произнесений одному и тому же диктору.

Для экспериментальных исследований использовалось несколько речевых баз:

— обучение универсальной фоновой модели (UBM) проводилось на базе NIST-2004 [3];

— тестовое множество составляли фонограммы 1348 дикторов-мужчин, взятые из базы NIST-2005;

— в качестве вспомогательной базы импостеров (самозванцев) для обучения SVM выбраны фонограммы 380 дикторов из базы Фишера [3].

Базовый эксперимент заключался в построении статистических моделей тестового произнесения и эталона на основе адаптации GMM-UBM-MAP [12] в виде супервекторов, полученных объединением средних компонент смеси гауссовых распределений, и последующей классификации в модуле SVN. Результаты базового эксперимента сравнивались с результатами экспериментов, в которых исследовалось влияние нелинейной редукции пространства супервекторов на эффективность системы верификации GMM-SVM.

Для построения системы распознавания дикторов с использованием методов обучения на основе многообразий (Isomap, Laplacian Eigenmaps) выполнялась следующая последовательность шагов.

1) Как и в базовом эксперименте, предварительно вычислялось N моделей (GMM-UBM) для эталонного, тестового произнесений и произнесений из базы SVM импостеров.

2) Супервекторы конфигурировались в виде матриц:

$$\mathrm{M}_1 = \begin{Bmatrix} \underset{\downarrow}{\overline{x}_1} & \underset{\downarrow}{\overline{x}_2} & \dots & \underset{\downarrow}{\overline{x}_C} \\ x(1,1) & x(1,2) & \dots & x(1,C) \\ \dots & \dots & \dots & \dots \\ x(D,1) & \dots & \dots & x(D,C) \end{Bmatrix}, \tag{1}$$

где $i \in 1, ..., N$, D — размерность векторов признаков, C — число компонент гауссовой смеси.

3) На основе полученных на предыдущем шаге матриц M_i, для каждого значения $d \in 1, ..., D$ формировались матрицы $\mathrm{A}^d_{N,C}$ размерности [N×C]:

$$\mathrm{A}^d_{N,C} = \begin{Bmatrix} \underset{\downarrow}{1} & \underset{\downarrow}{2} & \dots & \underset{\downarrow}{C} \\ M_1(d,1) & M_1(d,2) & \dots & M_1(d,C) \\ \dots & \dots & \dots & \dots \\ M_N(d,1) & \dots & \dots & M_N(d,C) \end{Bmatrix}. \tag{2}$$

Набор матриц $\mathrm{A}^d_{N,C}$ при $d \in 1, ..., D$ соответствует представлениям GMM моделей N произнесений в C-мерном пространстве.

4) Для каждого подмножества, выраженного матрицами $\mathrm{A}^d_{N,C}$ при $d \in 1, ..., D$, выполнялось снижение размерности пространства $\mathrm{R}^C \to \mathrm{R}^G$ с учетом сохранения топологических особенностей.

В качестве алгоритмов нелинейной редукции пространства использовались Isomap и Laplacian Eigenmaps:

$$F\left(\mathrm{A}^d_{N,C}\right) = \mathrm{A}^d_{N,G}, \tag{3}$$

F — алгоритм преобразования пространства, G — размерность нового пространства, $G << C$.

После выполнения редукции пространства на основе полученных матриц $A_{N,G}^{d}$ составлялись N моделей GMM, которые использовались для обучения SVM-классификатора и тестирования.

Было проведено два эксперимента с различными условиями. В первом сравнивались две схемы верификации диктора:

— базовая;

— схема с нелинейной редукцией пространства GMM супервекторов методом Isomap.

При реализации базовой схемы использовались смесь из 512 гауссовых компонент и 50-мерные векторы признаков, размер супервекторов для этого случая составил DC=25 600. Для редукции пространства во второй системе верификации применялся метод Isomap, с параметром k=12. Число гауссовых компонент, используемых для описания супервекторов GMM после редукции, составило G=128, что соответствовало DG=6400.

Результаты первого эксперимента показали, что нелинейный метод редукции пространства может быть использован для сокращения размерности GMM-моделей без потери качества верификации. При сокращении размерности моделей в 4 раза по сравнению с базовой схемой уровень ошибки верификации не изменился и остался на уровне EER≈7,5 %.

Во втором эксперименте выполнялась редукция числа компонент двумя методами: методом Laplacian Eigenmaps и методом главных компонент (МГК). Базовая схема была реализована с использованием C=128 гауссовых компонент при размерности пространства векторов признаков D=50, что соответствовало размерности супервекторов DC=6400. В результате линейной редукции методом МГК число компонент для описания моделей было сокращено с 128 до 64, соответственно DC=3200.

Размерность нового пространства при нелинейной редукции методом Laplacian Eigenmaps с использованием 12 ближайших соседей была выбрана такой же, как в случае МГК, чтобы сравнить эффективность линейного и нелинейного методов при одинаковых параметрах.

Результаты экспериментов показали, что при сокращении размерности моделей всего в 2 раза уровень равновероятной ошибки первого и второго рода базовой схемы верификации EER=7,65 %, после редукции методами МГК и Laplacian Eigenmaps ухудшился соответственно до EER=8,53 и 7,95 %. Отсюда можно сделать вывод, что для верификации диктора более эффективно проводить редукцию пространства нелинейными методами. В этом случае учет топологии пространства распределения речевых данных позволяет более точно определить признаки, соответствующие индивидуальным особенностям дикторов.

Заключение. Несмотря на то что методы Isomap, Laplacian Eigenmap позволяют эффективно выполнить нелинейную редукцию пространства с учетом топологической структуры данных, целесообразность их применения к задачам, связанным с распознаванием голоса, неочевидна. Основная причина этого связана с тем, что для отображения тестовых данных в пространство пониженной размерности необходимо установить их связи с обучающими данными. Это сопряжено со значительными вычислительными затратами, поскольку подразумевает определение ближайших соседей тестового образца, определение структуры в виде взвешенных графов, а затем требует редукции пространства.

В работе [13] отмечается, что методы Isomap, Laplacian Eigenmap эффективны для снижения размерности пространства, но они не обеспечивают оптимальных условий для классификации после преобразования данных. Поэтому помимо нелинейного преобразования пространства требуется увеличить дискриминативность признаков, например, дополнительно используя линейный дискриминативный анализатор Фишера или применяя анализ главных компонент [13, 14].

Построение графов в методе многообразий, которое сводится к определению ближайших соседей в каждой точке множества, является важным этапом для определения структуры данных, от которого зависит качество классификации. В работе [14] приведен способ определения ближайших соседей, при котором достигается оптимальное для классификации построение графов.

Использование метода многообразий на текущем этапе развития систем текстонезависимого распознавания диктора не распространено. Это связано с недоказанностью его эффективности (повышением надежности распознавания и достаточным уменьшением размерности пространства признаков, по сравнению с другими современными методами) при решении задач идентификации и верификации дикторов, а также их большой вычислительной сложностью.

Согласно результатам международных конкурсов NIST за последние несколько лет [3, 4, 15, 16], основными методами редукции размерности, доказавшими свою эффективность, являются совместный факторный анализ (Joint Factor Analysis, JFA) [17], метод полной изменчивости (Total Variability, TV) [18], вероятностный линейный дискриминантный анализ (Probabilistic Linear Discriminative Analysis, PLDA) [18]. Эти методы оперируют моделями статистических распределений данных и не учитывают особенности их структуры на локальном уровне. Дальнейшее развитие этих методов для распознавания дикторов, по всей видимости, будет связано с усложнением статистических моделей и привлечением вариационного байесовского анализа для определения параметров распределений [18].

Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

СПИСОК ЛИТЕРАТУРЫ

1. *Cayton L.* Algorithms for manifold learning. UCSD tech report CS2008-0923. University of California, San Diego, 2005. 17 p.
2. *Матвеев Ю. Н.* Технологии биометрической идентификации личности по голосу и другим модальностям // Вестн. МГТУ им. Н. Э. Баумана. Сер. „Приборостроение“. 2012. № 3 „Биометрические технологии“. С. 46—61.
3. *Kozlov A., Kudashev O., Matveev Yu., Pekhovsky T., Simonchik K., Shulipa A.* SVID Speaker Recognition System for NIST SRE 2012 // Proc. of 15th Intern. Conf. “Speech and Computer” (SPECOM 2013). Springer Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence. 2013. Vol. 8113. P. 278—285.
4. *Матвеев Ю. Н., Симончик К. К.* Система идентификации дикторов по голосу для конкурса NIST SRE 2010 // Тр. 20-й Междунар. конф. по компьютерной графике и зрению ГрафиКон’2010. СПб: СПбГУ ИТМО, 2010. С. 315—319.
5. *Michalevsky Y., Talmon R., Cohen I.* Speaker Identification Using Diffusion Maps // Proc. 19th Europ. Signal Processing Conf. (EUSIPCO-2011). Barcelona, Spain, 2011. P. 1299—1302.
6. *Sierra G. H., Bonastre J.-F., Matrouf D., Calvo J. R.* Topological representation of speech for speaker recognition // Proc. INTERSPEECH-2010. 2010. P. 2134—2137.
7. *Lafon S. S.* Diffusion maps and geometric harmonics: PhD thesis. Yale University, 2004.
8. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 47—51.
9. *Davis S., Mermelstein P.* Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE Transact. Acoustics, Speech and Signal Processing. 2002. Vol. 28, N 4. P. 357—366.
10. *Campbell J. P., Jr.* Testing with the YOHO CD-ROM voice verification corpus // Proc. IEEE Intern. Conf. Acoust., Speech Signal Process. 1995. Vol. 1. P. 341—344.

11. *Bimbot F., Bonastre J.-F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacretaz D., Reynolds D.* A tutorial on text independent speaker verification // EURASIP J. Appl. Signal Process. 2004. Vol. 4. P. 430—451.

12. *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.

13. *Yang M.-H.* Extended Isomap for Pattern Classification // Proc. 18th National Conf. on Artificial Intelligence. 2002. P. 224—229.

14. *Wu Y., Chan K., Wang L.* Face Recognition based on Discriminative Manifold Learning // Proc. IEEE Intern. Conf. on Pattern Recognition. 2004. Vol. 4. P. 171—174.

15. *Burget L., Fapso M., Hubeika V., Glembek O., Karafiat M., Kockmann M., Matejka P.* BUT system for NIST 2008 speaker recognition evaluation // Proc. Interspeech. 2009. P. 2335—2338.

16. Loquendo - Politecnico Di Torino's 2010 NIST Speaker Recognition Evaluation System // Proc. ICASSP. 2011. P. 5464—5467.

17. *Kenny P.* Joint factor analysis of speaker and session variability: Theory and algorithms. Tech. Report CRIM-06/08-13. 2005.

18. *Dehak N., Dehak R., Kenny P., Brummer N., Ouellet P., Dumouchel P.* Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification // Proc. Interspeech. 2009. P. 1559—1562.

19. *Kenny P.* Bayesian Speaker Verification with Heavy-Tailed Priors // Proc. Odyssey Speaker and Language Recognition Workshop. Brno, Czech Republic, 2010. P. 1—10.

Сведения об авторах

Юрий Николаевич Матвеев — д-р техн. наук, профессор; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ-инновации", Санкт-Петербург; главный научный сотрудник; E-mail: matveev@mail.ifmo.ru

Андрей Константинович Шулипа — ООО „ЦРТ-инновации", Санкт-Петербург; научный сотрудник; E-mail: shulipa@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.