УДК 621.391.037.372

СИСТЕМА ИДЕНТИФИКАЦИИ ВОЗРАСТНОЙ ГРУППЫ ГОВОРЯЩЕГО ПО ЗАПИСЯМ СПОНТАННОЙ РЕЧИ

К.К. Симончик

Предлагается использовать популярный в текстонезависимой идентификации диктора метод выделения **i**-векторов для решения задачи идентификации возрастной группы говорящего. Исследуется две реализации системы идентификации возрастной группы говорящего: предложен подход на базе машины опорных векторов, а также подход на основе линейной регрессионной модели. В обоих случаях была достигнута хорошая надежность детектирования возрастной группы диктора по записям фонограмм устной речи. Средний процент правильной идентификации возрастной группы диктора составил 61% и 65% соответственно на речевой базе NIST SRE 2008.

Ключевые слова: возраст, **i**-вектор, SVM, линейная регрессия.

Введение

В современном мире широко развиваются речевые технологии различного применения – синтеза, идентификации, распознавания речи. Последние две технологии позволяют выделить из речи информацию различного типа – «кто» говорит на фонограмме и «что» именно говорится. Однако по записи речи можно получить и иную информацию – пол диктора, его эмоциональное состояние и т.д. Работа посвящена проблеме определения возраста диктора по записям его устной речи.

Информация о возрасте диктора важна для применения в различных областях. Например, в зависимости от возраста говорящего могут применяться соответствующие акустические модели для распознавания речи. Контроль возраста ребенка позволит разграничить доступ к информационному контенту Интернета. Автоматизированная сегментация голосовых запросов в колл-центр по возрасту позволит сделать их обработку эффективнее и т.д.

На сегодняшний день наиболее популярными подходами в определении возрастной группы говорящего являются модели на основе использования смесей гауссовых распределений (СГР) и машин опорных векторов (support vector machine, SVM) [1–4], а также скрытых марковских моделей [4]. В настоящей работе для решения задачи идентификации возрастной группы предложено использовать широко применяемый в текстонезависимой идентификации по голосу подход, основанный на представлении модели речи диктора в виде многомерного вектора, который используется в качестве входных данных для классификатора возрастной группы. Проводится исследование двух методов классификации: с использованием SVM и на основе линейной регрессионной модели.

Описание схемы выделения признаков

Схема выделения **i**-векторов в качестве признаков подразумевает анализ всего речевого материала, присутствующего на фонограмме, и его последовательное преобразование в многомерный вектор (рис. 1). Более подробно об этом можно узнать в [6].

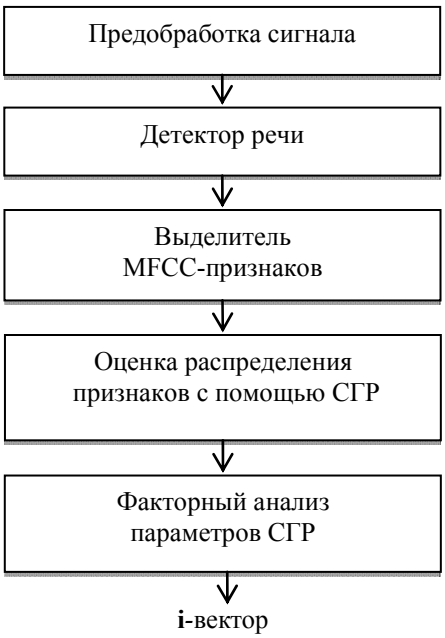

Рис. 1. Схема выделителя признаков

Перед модулем выделения речи использовались специальные алгоритмы предобработки всего сигнала [7–9]. Так как на реальных записях, сделанных в обычных офисных или бытовых условиях, часто присутствуют посторонние сигналы, такие как импульсные, мультитональные помехи, а также музыка и перегруженные участки речи, предварительная отбраковка непригодных для анализа участков фонограммы позволяет повысить эффективность дальнейшей обработки детектором речи.

Далее использовался детектор речи, основанный на оценке энергии сигнала в речевом диапазоне частот.

В качестве речевых признаков использовались мэл-частотные кепстральные коэффициенты (Mel frequency cepstral coefficients, MFCC) размерностью 39 [10, 11]. Длина временного окна установлена в 20 мс с шагом 10 мс. Распределение MFCC-признаков моделировалось с помощью СГР [10]. В качестве универсальной фоновой модели (Universal background model, UBM) была использована СГР размерностью 512 смесей. База обучения UBM состояла из порядка 60000 фонограмм (более 4000 дикторов), взятых из речевых баз NIST SRE 98, 99, 2000, 2003, 2004, 2005, 2006, 2008 (мужской и женский гендеры). Далее на основе подхода, описанного в [1], производился факторный анализ параметров СГР с целью понижения размерности данных и выделялся скрытый **i**-вектор:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\boldsymbol{\omega} + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\mu}$ – супервектор параметров СГР модели диктора; $\mathbf{m}$ – супервектор параметров UBM; $\mathbf{T}$ – матрица, задающая базис в редуцированном пространстве признаков; $\boldsymbol{\omega}$ – **i**-вектор в редуцированном пространстве признаков, $\boldsymbol{\omega} \in N(0,1)$; $\boldsymbol{\varepsilon}$ – вектор ошибки.

Использование классификатора на основе SVM

SVM является классификатором, впервые предложенным Вапником [11]. Идея SVM заключается в поиске оптимальной разделяющей гиперплоскости между двумя классами. Формула, задающая решающую функцию SVM, приведена ниже:

$$f(\mathbf{x}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x_i}) + w_0,$$

где N – число опорных векторов; $\mathbf{x_i}$ – **i** -й опорный вектор; $K(.,.)$ – ядро SVM. Так как SVM предназначен для разделения множества входных данных на 2 класса, а возрастных групп в исследуемой задаче было более двух, то была использована стратегия «турнир на выбывание» [12], которая предполагала обучение $M(M–1)/2$ классификаторов. Каждый SVM-классификатор был обучен на разделение только двух групп, а окончательное решение о принадлежности **i**-вектора к той или иной группе выносилось по анализу решений всех классификаторов.

Для принятия решения о принадлежности объекта возрастной группе использовалась следующая логика. На каждом шаге распознавание **i**-вектора осуществлял только один классификатор – «победившая» группа продолжала борьбу и определяла следующий используемый классификатор. Процесс осуществлялся до тех пор, пока не оставалась одна победившая группа, которая считалась результатом распознавания.

Таким образом, схема детектора возраста с использованием SVM-классификатора выглядела, как показано на рис. 2. Было выбрано линейное ядро SVM как достаточно простое и наиболее робастное.

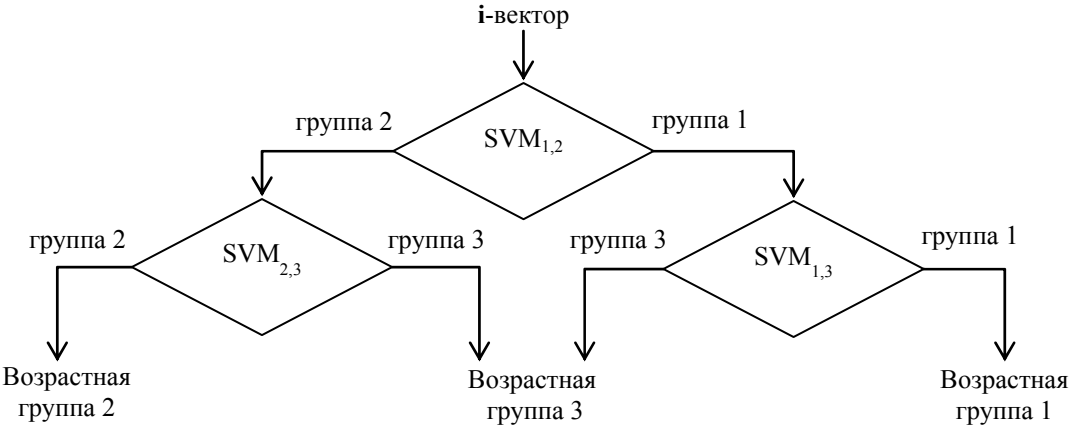

Рис. 2. Блок-схема детектора возраста на основе SVM

Классификатор на основе линейной регрессионной модели

Модель линейной регрессии позволяет определить линейную зависимость между скаляром y и несколькими наблюдаемыми переменными [14]. Линейная регрессия задается следующим уравнением:

$$y = \mathbf{A}\boldsymbol{\omega} + b,$$

где $\boldsymbol{\omega}$ – наблюдаемый вектор признаков (**i**-вектор); $\mathbf{A}$ – матрица обобщенной линейной модели; b – смещение; y – выходная переменная, связанная с возрастом диктора. Оптимизация параметров модели производилась с использованием минимизации функции квадратичного отклонения:

$$\sum_{i=1}^{M} (y - \hat{y})^2 \to \min,$$

где M – объем выборки. Далее производилось разбиение выходного скаляра y на возрастные классы с использованием простейшего порогового классификатора (рис. 3).

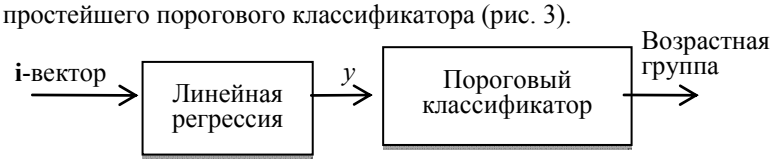

Рис. 3. Блок-схема детектора возраста на основе линейной регрессии

Эксперименты

Для экспериментов была взята речевая база NIST SRE 2008. Из нее были отобраны фонограммы тех дикторов, возраст которых можно было рассчитать. Оценка возраста производилась как разница между датой записи и годом рождения диктора. Всего было отобрано 16077 фонограмм 439 дикторов мужчин и 727 дикторов женщин возраста от 17 до 84 лет, говорящих на английском, китайском, арабском, русском и других языках (всего более 10 языков). Каналы связи были представлены микрофоном и телефоном. Речевая база была разбита на два равных множества, Train (обучающее) и Test (тестовое), таким образом, чтобы дикторы из разных множеств не пересекались.

Фонограммы дикторов были поделены на 3 возрастные группы:

1. от 16 до 25 лет –Young Adult (Молодые люди);
2. от 26 до 70 лет – Adult (Взрослые);
3. старше 70 лет – Senior (Пожилые).

На обучающем материале Train было рассчитано 3 SVM-классификатора для разделения попарно следующих возрастных групп: Young Adult/Adult, Young Adult/Senior и Adult/Senior. На множестве Train также был обучен второй классификатор, базирующийся на модели линейной регрессии. На рис. 4 показана зависимость выходного значения y от реального возраста диктора, записанного на фонограмме.

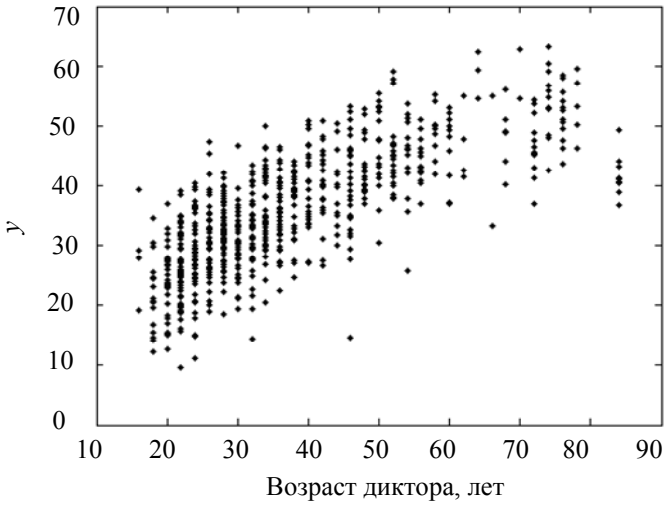

Рис. 4. Зависимость параметра *y* от возраста диктора на фонограмме

На рисунке видно, что существует достаточно высокая корреляция между параметром *y* регрессионной модели и возрастом диктора. Тем не менее, зависимость нельзя назвать строго линейной, что делает уместным использование индивидуальных пороговых значений *y* для каждой возрастной группы.

Далее для каждой из двух систем было проведено тестирование на множестве Test. В табл. 1, 2 показаны матрицы результатов классификации с помощью SVM-классификатора и классификатора на базе

линейной регрессионной модели. На рис. 5 показаны распределения выхода второго классификатора для различных возрастных групп. Стоит отметить, что распределения групп, не являющихся соседними, Young Adult и Senior, почти не пересекаются.

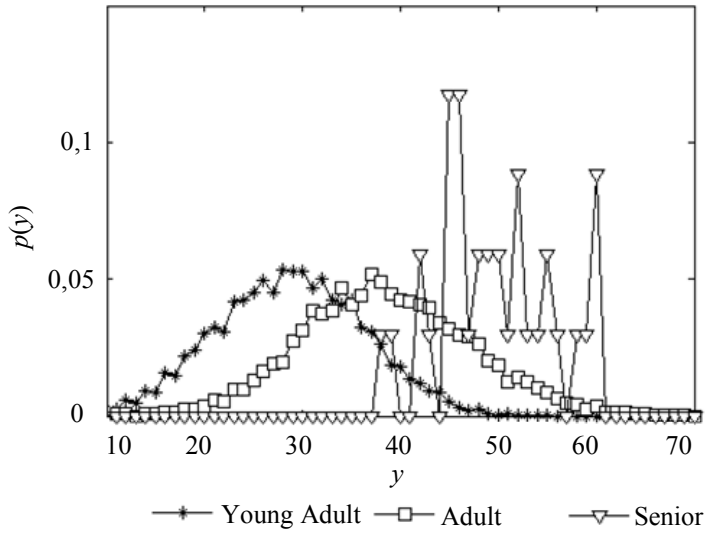

Рис. 5. Распределение параметра *y* регрессионной модели для каждой возрастной группы

Тест	Результаты классификации		
	Young Adult	Adult	Senior
Young Adult	**72%**	27%	1%
Adult	26%	**72%**	2%
Senior	0%	62%	**38%**

Таблица 1. Результаты экспериментов системы на основе SVM классификатора

Тест	Результаты классификации		
	Young Adult	Adult	Senior
Young Adult	**64%**	35%	1%
Adult	21%	**58%**	21%
Senior	0%	26%	**74%**

Таблица 2. Результаты экспериментов системы на основе линейной регрессионной модели

Заключение

В работе представлено два подхода к реализации детектора возраста говорящего на основе представления модели голоса в виде **i**-вектора.

В первую очередь необходимо заметить, что основная ошибка детектирования наблюдается только между смежными возрастными группами (Young Adult/Adult and Adult/Senior), тогда как ошибка между группами Young Adult и Senior не превышает 1%.

SVM-классификатор показывает хорошую надежность классификации для групп Young Adult и Adult, однако для группы Senior его точность в 2 раза ниже, чем у подхода на основе линейной регрессионной модели. Этот результат может быть объяснен недостаточным объемом обучающего множества группы Senior для SVM-классификатора. В отличие от SVM, линейная регрессионная модель демонстрирует более стабильный результат по всем возрастным группам и, в конечном итоге, показывает более высокую среднюю надежность детектирования возраста (65% против 61% у SVM).

Линейная регрессионная модель, обеспечивающая проекцию параметра *y*, одинаковую для всех групп, обеспечивает лучшую надежность и робастность идентификации возраста для набора данных, приведенных в данной работе.

В целом можно констатировать, что средняя надежность идентификации указанных возрастных групп по обоим представленным методам оказалась в 1,5 раза выше, чем у подхода, описанного в работе [4].

Литература

1. Feld M., Burkhardt F., Muller C. Automatic Speaker Age and Gender Recognition in the Car for Tailoring Dialog and Mobile Services // INTERSPEECH-2010. – 2010. – P. 2834–2837.

2. Bocklet T., Maier A., Bauer J.G., Burkhardt F., Noth E. Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines // Proceedings of ICASSP 2008. – Las Vegas, 2008. – P. 1605–1608.
3. Dobry G., Hecht R.M., Avigal M., Zigel Y. Dimension Reduction Approaches for SVM based Speaker Age Estimation // INTERSPEECH-2009. – Brighton, UK, 2009. – P. 2031–2034.
4. Porat R., Lange D., Zigel Y. Age recognition based on speech signals using weights supervector // INTERSPEECH-2010. – 2010. – P. 2814–2817.
5. Metze F., Ajmera J., Englert R., Bub U., Burkhardt F., Stegmann J., Müller C., Huber R., Andrassy B., Bauer J.G., Little B. Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications // Proceedings of ICASSP 2007. – Honolulu, 2007. – V. 4. – P. 1089–1092.
6. 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 // INTERSPEECH-2009. – Brighton, UK, 2009. – P. 1559–1562.
7. Алейник С.В., Матвеев Ю.Н., Раев А.Н. Метод оценки уровня клиппирования речевого сигнала // Научно-технический вестник информационных технологий, механики и оптики. – 2012. – № 3 (79). – С. 79–83.
8. Козлов А.В., Лоханова А.И., Симончик К.К. Алгоритм детектирования музыкальных фрагментов в задачах речевой обработки // Научно-технические ведомости СПбГПУ. – 2010. – № 4 (103). – С. 7–11.
9. Симончик К.К., Галинина О.С., Капустин А.И. Алгоритм обнаружения речевой активности на основе статистик основного тона в задаче распознавания диктора // Научно-технические ведомости СПбГПУ. – 2010. – № 4 (103). – С. 18–23.
10. Белых И.Н., Капустин А.И., Козлов А.В., Лоханова А.И., Матвеев Ю.Н., Пеховский Т.С., Симончик К.К., Шулипа А.К. Система идентификации дикторов по голосу для конкурса NIST SRE 2010 // Информатика и ее применения. – 2012. – Т. 6. – Вып. 1. – С. 24–31.
11. Spiegel W., Stemmer G., Lasarcyk E., Kolhatkar V., Cassidy A., Potard B., Shum S., Song Y.C., Xu P., Beyerlein P., Harnsberger J., Nöthm E. Analyzing Features for Automatic Age Estimation on Cross-Sectional Data // INTERSPEECH-2009. – Brighton, UK. – 2009. – P. 2923–2926.
12. Vapnik V.N. Statistical Learning Theory. – New York: Wiley, 1998. – 732 p.
13. Platt J., Cristianini N., Shawe-Taylor J. Large Margin DAGS for Multiclass Classification // Advances in Neural Information Processing Systems. – MIT Press, 2000. – P. 547–553.
14. Rawlings J., Pantula S. Dickey D. eds. Applied Regression Analysis. – Springer Texts in Statistics, 1998. – 736 p.

Симончик Константин Константинович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, ООО «ЦРТ-инновации», кандидат технических наук, доцент, руководитель отдела, simonchik@speechpro.com