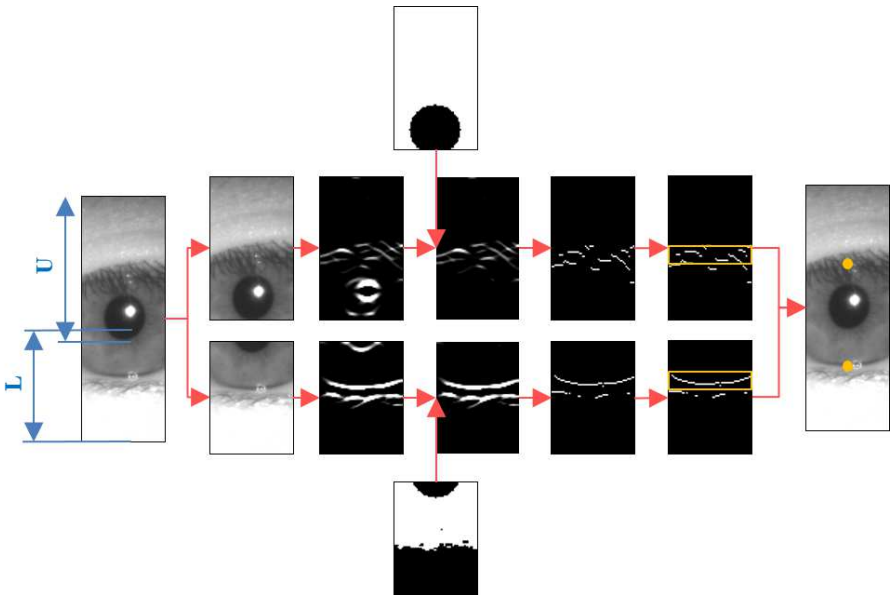
ISSN 2223-3792

Машинное обучение и анализ данных

2016 год **Том 2, номер 4**



Машинное обучение и анализ данных

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретических основ информатики и приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на английском и русском языках.

Журнал включен в российский индекс научного цитирования РИНЦ. Информация о цитировании статей находится на сайте Российского индекса научного цитирования, ISSN 2223-3792, номер свидетельства о регистрации ЭЛ № ФС 77-55486. Журнал зарегистрирован в системе Crossref, doi `http://dx.doi.org/10.21469/22233792`.

- Новостной сайт http://jmlda.org/
- Электронная система подачи статей http://jmlda.org/papers/
- Правила подготовки статей http://jmlda.org/papers/doc/authors-guide.pdf

Тематика журнала:
— классификация, кластеризация, регрессионный анализ;
— алгебраический подход к проблеме синтеза корректных алгоритмов;
— многомерный статистический анализ;
— выбор моделей и сложность;
— глубокое обучение;
— статистическая теория обучения;
— методы прогнозирования временных рядов;
— методы обработки и распознавания сигналов;
— методы оптимизации в задачах машинного обучения и анализа данных;
— методы анализа больших данных;
— методы визуализации данных;
— обработка и распознавание речи и изображений;
— анализ и понимание текста;
— информационный поиск;
— прикладные задачи анализа данных.

Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2016

Journal of Machine Learning and Data Analysis

The journal Machine Learning and Data Analysis publishes original research papers and reviews of the developments in the field of artificial intelligence, theoretical computer science and its applications. The journal aims to promote the theory of machine learning and data mining and methods of conducting computational experiments. Papers are accepted in English and Russian.

The journal is included in the Russian science citation index RSCI. Information about citation to articles can be found at the Russian science citation index website. ISSN 2223-3792. Mass media registration certificate ЭЛ № ФС 77-55486. The Crossref journal doi is http://dx.doi.org/10.21469/22233792.

- Journal news and archive http://jmlda.org/
- Open journal system for papers submission http://jmlda.org/papers/
- Style guide for authors http://jmlda.org/papers/doc/authors-guide.pdf

The scope of the journal:
— classification, clustering, regression analysis;
— multidimensional statistical analysis;
— Bayesian methods for regression and classification;
— model selection and complexity;
— deep learning;
— Statistical Learning Theory;
— time series forecasting techniques;
— methods of signal processing and speech recognition;
— optimization methods for solving machine learning and data mining problems;
— methods of big data analysis;
— data visualization techniques;
— methods of image processing and recognition;
— text analysis, text mining and information retrieval;
— applied data analysis problems.

Dorodnicyn Computing Centre FRC CSC RAS
Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics
Division "Intelligent Systems"

Moscow, 2016

Содержание

В. Я. Чучупал
Неявная модель вариативности произношения для автоматического распознавания речи 370

А. А. Жарких, А. В. Горбунов
Реализация пакета программ для встраивания и извлечения скрытых сообщений в аудиофайлах 378

И. В. Бахмутова, В. Д. Гусев, Л. А. Мирошниченко, Т. Н. Титкова
Сопоставление и интеграция подходов к дешифровке древнерусских знаменных песнопений 391

Н. А. Волков, М. Е. Жуковский
Вероятностная модель для сглаживания целевых метрик качества ранжирования 407

Н. Д. Смелик, А. А. Фильченков
Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления 421

Г. А. Одиноких, В. С. Гнатюк, М. В. Коробкин, В. А. Еремеев
Метод определения положения век на изображении при распознавании человека по радужной оболочке глаза с мобильного устройства 442

В. П. Кальян
Выбор решений при распознавании эмоций по речи 454

Contents

V. J. Chuchupal
Implicit pronunciation variation model for automatic speech recognition 370
A. A. Zharkikh and A. V. Gorbunov
Implementation of the software package for embedding and extracting hidden messages in audiofiles 378
I. V. Bakhmutova, V. D. Gusev, L. A. Miroshnichenko, and T. N. Titkova
Comparison and integration of approaches to deciphering Russian ancient chants . . . 391
N. A. Volkov and M. E. Zhukovskii
On a probabilistic model for smoothing discrete ranking quality metrics 407
N. D. Smelik and A. A. Filchenkov
Multimodal topic model for texts and images utilizing their embeddings 421
G. A. Odinokikh, V. S. Gnatyuk, M. V. Korobkin, and V. A. Eremeev
Method of eyelid detection on image for mobile iris recognition 442
V. P. Kalyan
Decision support in process of recognizing emotion in speech 454

Неявная модель вариативности произношения для автоматического распознавания речи*

В. Я. Чучупал
v.chuchupal@gmail.com
ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Вариативность произношения слов и словосочетаний в естественной разговорной речи является одним из основных источников ошибок при автоматическом распознавании речи, поэтому использование моделей вариативности произношения представляется важным направлением повышения эффективности работы систем распознавания речи. Рассматривается проблема моделирования вариативности, которая вызвана нечеткой, неполной артикуляцией, например в результате нарушения синхронизации работы органов речеобразования. Предлагается использовать неявные произносительные модели, основанные на комбинировании акустических моделей соседних звуков. Комбинирование может заключаться в сглаживании или интерполяции параметров акустических моделей текущих звуков параметрами соседних моделей. Степень проявления вариативности, вообще говоря, зависит от синтаксического и просодического контекста звука, поэтому предлагается использовать меняющиеся параметры интерполяции в зависимости от наличия позиционных, фонетических, синтаксических и просодических признаков. Подход с комбинированием акустических моделей на основе сглаживания их параметров был описан в научной литературе, однако автору неизвестны исследования с комбинированием именно соседних акустических моделей, где бы параметры комбинирования зависели от текущих контекстных признаков. Предварительные эксперименты на корпусах с читаемой и разговорной речью показали справедливость предположений о целесообразности использования интерполяционных моделей и существования зависимости параметров сглаживания моделей от наличия позиционных и синтаксических признаков.

Ключевые слова: *автоматическое распознавание речи; вариативность речи; моделирование произношения; скрытые марковские модели*

DOI: 10.21469/22233792.2.4.01

1 Введение

Использование моделей вариативности произношения имеет высокий потенциал как способ повышения эффективности автоматического распознавания речи. Это подтверждается данными так называемых симуляционных экспериментов, когда за счет использования корректных фактических произносительных транскрипций уровень пословной ошибки распознавания — WER (word error rate [1]) — понижался в разы [2, 3].

В литературе встречаются два основных подхода к моделированию вариативности произношения. Явное моделирование (explicit modeling) заключается в моделировании вариативности произнесения путем описания возможных изменений в фонемной транскрипции слов [4]. Неявное моделирование (implicit modeling) [5] описывает вариативность произнесения путем изменений в структуре моделей звуков в канонической транскрипции слов. В прикладных системах обычно используется явное моделирование, которое естественным

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-01-00607.

образом описывается в рамках классической статистической формулировки распознавания слитной речи. Если $X = \{x_t\}, t = 1, \ldots, T$, — наблюдаемый образ в виде последовательности параметров речевого сигнала, а $W = \{w_i\}, i = 1, \ldots, N$, — последовательность слов словаря, то результат распознавания X в виде наиболее вероятной последовательности слов W^* определятся из уравнения [6]:

$$W^* = \arg\max_W P(W|X) = \arg\max_W \frac{P(X|W)P(W)}{P(X)} = \arg\max_W P(X|W)P(W). \qquad (1)$$

Первый сомножитель $P(X|W)$ в числителе (1) соответствует правдоподобию данных при заданной последовательности слов и определяется с помощью акустических моделей. Полученная величина правдоподобия затем умножается на значение $P(W)$, которое определяется с помощью модели языка. Знаменатель $P(X)$ — вероятность наблюдения X, выполняет функции нормализующего члена.

Обозначим фонемную транскрипцию слова w через t^w, множество всех фонемных транскрипций этого слова обозначим T^w. Множество возможных транскрипций последовательности слов W обозначим T^W. Запись t^W будет использоваться для обозначения какой-либо конкретной последовательности транскрипций из T^W. Тогда уравнение (1) можно аппроксимировать (так называемая аппроксимация Витерби) выражением:

$$W^* \approx \arg\max_{W, t^W \in T^W} P(X|t^W)P(t^W|W)P(W).$$

Оценка $P(t^W|W)$ осуществляется моделью вариативности произношения, параметрами которой являются фонемные транскрипции T^W и условные вероятности их реализации $\{P(t^W|W), t^W \in T^W\}$.

Таким образом, явные модели вариативности можно идентифицировать на основе используемых методов выбора фактических фонемных транскрипций и определения их вероятностей. Сложность заключается в том, что наиболее очевидный способ выбора фактических транскрипций с помощью фонемного распознавателя до недавнего времени был неэффективен ввиду низкой точности таких распознавателей, а использование естественных частотных оценок для вероятностей реализации транскрипций затруднительно по причине отсутствия корпусов данных требуемого размера.

Различные способы преодоления этих проблем достаточно широко описаны в литературе, но полученный за счет использования явных моделей выигрыш в величине WER для разговорной речи не так велик, как можно было ожидать: 0,8% [4, 7], 2,2% [8], 1,8% [9], 0,9% [10].

Основой явных моделей вариативности является предположение, что произносительные изменения в разговорной речи можно достаточно адекватно описать полными заменами (включая вставку и удаление) фонем. В то же время анализ экспериментальных данных показывает [5], что более точным описанием вариативности произношения, особенно в типичной ситуации, когда вариативность вызвана нарушением синхронизации движений речевых органов [11], является использование моделей, способных представлять неполные изменения фонемного качества звуков. Такие произносительные модели относят к так называемым неявным.

В отличие от явных моделей неявные модели реализуются как часть акустических моделей, например за счет усложнения их структуры либо использования множественных моделей. Практически выигрыш в точности распознавания для неявных моделей не отличается существенно от такового же для явных в терминах величины пословной ошибки:

1,7% [5], 0,7% [12, 13], 2,2% [14], 2,39% (послоговой ошибки для китайского языка) [15], 2,5% [16].

Несмотря на наличие исследований в пользу условного характера проявления вариативности в разговорной речи [17], которые можно интерпретировать как возможность предсказать появление вариативности исходя из синтаксических и семантических характеристик речевого сигнала, в литературе мало конкретных результатов в этом направлении.

2 Моделирование вариативности произношения путем сглаживания параметров акустических моделей

Пусть m и n обозначают звуки, а $P(x|m), P(x|n)$ — их акустические модели, которые определяют условные вероятности наблюдения параметров x. Интерполированную (или сглаженную) модель для m, $P(x|m,n)$ определим как [15]:

$$P_\lambda(x|m,n) = \lambda P(x|m) + (1-\lambda)P(x|n), 0 \leqslant \lambda \leqslant 1. \tag{2}$$

Форма (2) позволяет упрощенно описать некоторые частые эффекты вариативности произношения. Например, значение коэффициента $\lambda = 0$ означает, что звук m пропущен, а $\lambda = 0{,}5$ соответствует частичному изменению его качества, например оглушению, озвончению, назализации, если n обладает этими признаками.

Если реализация звука m соответствует параметрам $x_{s(m)}, \ldots, x_{e(m)}$ на отрезке времени $s(m), \ldots, e(m)$, то наиболее правдоподобная оценка коэффициента λ для (2) вычисляется аналогично оценке весов смесей при обучении GMM (gaussian mixture models) моделей [18]:

$$\lambda_{m,n} = \frac{\sum_{t=s(m)}^{e(m)} P(x_t|m)}{\sum_{t=s(m)}^{e(m)} (P(x_t|m) + P(x_t|n))}. \tag{3}$$

Для корпуса данных из R реализаций (предложений) $U = \{u_r | r = 1, \ldots, R\}$ с, вообще говоря, несколькими произносительными вариантами, транскрипциями $\{f_r | r = 1, \ldots, F_r\}$, значение параметра $\hat{\lambda}_{m,n}$ можно вычислить, усредняя локальные значения (3) по всем вхождениям пар звуков (m,n):

$$\hat{\lambda}_{(m,n)} = \frac{\sum_{r=1}^{R} \sum_{f=1}^{F_r} \sum_{(m,n)} \lambda_{(m,n)} P(m)}{\sum_{r=1}^{R} \sum_{f=1}^{F_r} \sum_{(m,n)} P(m)}, \tag{4}$$

где $P(m)$ — правдоподобие звука m для вхождения (m,n).

Поскольку модель (2) не использует никакой информации, кроме параметров соседних (в экспериментах n была правым контекстом m) моделей, учитывая результаты [15], можно ожидать, что заметного выигрыша от ее использования не будет.

Определим вектор признаков вариативности как вектор V, составленный из контекстных признаков, которые вероятно коррелируют с проявлениями вариативности произношения [17]: $V(c, l, r, \mathrm{nPh}, \mathrm{pPOS}, \mathrm{ROS}, \mathrm{wPOS}, \mathrm{POS}, \mathrm{mWrd}, \mathrm{fWrd}, \mathrm{LM})$, где

$$\left.\begin{array}{rl} c: & \text{центр;} \\ l: & \text{левый контекст;} \\ r: & \text{правый контекст;} \\ \text{nPh}: & \text{следующая фонема;} \\ \text{pPOS}: & \text{позиция фонемы;} \\ \text{ROS}: & \text{темп речи;} \\ \text{wPOS}: & \text{позиция слова;} \\ \text{POS}: & \text{часть речи слова;} \\ \text{mWrd}: & \text{словосочетание;} \\ \text{fWrd}: & \text{частотность слова;} \\ \text{LM}: & \text{значение модели языка.} \end{array}\right\} \quad (5)$$

Проверим предположение о зависимости степени вариативности произношения от наличия признаков из набора (5) и, в случае положительного ответа, построения интерполяционной модели в форме (2), которая может быть использована при распознавании речи.

3 Экспериментальное подтверждение эффективности модели вариативности произношения

Проверка утверждения, что использование интерполированных моделей вариативности произношения эффективно с точки зрения повышения точности автоматического распознавания речи очевидно должна проводиться в рамках экспериментов по распознаванию. Проведение таких экспериментов требует встраивания интерполированных моделей в существующий программный код для оценки параметров и распознавания, существенной модификации программного обеспечения. На данном этапе работы проверка эффективности осуществляется косвенными методами, т. е. путем вычисления и анализа значений параметра интерполяции λ на тестовых данных.

Для проведения экспериментальных исследований используется материал корпусов данных для русского языка: TeCoRus [19], RuSpeech [20] и PronExRu [21].

Данные разделены на три части: обучающую, настроечную и тестовую выборки. Обучающая выборка, на которой оценивались параметры акустических моделей, включала материал корпусов RuSpeech и TeCoRus (обучающую выборку), в основном, читаемую речь от 200+ чел. Настроечная выборка, использованная для оценки параметра λ, состояла из тестового материала корпуса TeCoRus (1000 предложений от 10 чел.). Тестовая включает материал корпуса PronExRu.

На обучающих данных были построены акустические модели звуков, контекстно-зависимые скрытые марковские модели трифонов, из трех состояний, гендер-зависимые. Всего обучались параметры около 10 000 состояний. Оценка признаков (5), за исключением темпа речи ROS, делается на основе данных из произносительного словаря, пополненного информацией о части речи слов.

Вычисление параметра ROS осуществляется на основе так называемого относительного темпа речи (relative ROS, [14]).

В ходе предварительных экспериментов для каждого центрального состояния акустических моделей рассчитывались бинарные значения признаков из (5), а по (4) вычислялись значения параметра λ интерполированных моделей.

Для ранжирования и последующего выбора значений λ по признакам вариативности строилось бинарное дерево решений. Вопросы для формирования дерева относились к наличию или отсутствию соответствующих признаков вариативности, например, «принадлежит ли звук, содержащий моделируемое состояние функциональному слову?», «принадлежит ли звук окончанию слова?» и т. п. В качестве критерия для выбора лучшей пары «вопрос–лист» дерева для разбиения на текущем шаге алгоритма использовалось изменение значения энтропии параметра λ в результате расщепления листа.

Полученные результаты имеют предварительный характер. Можно утверждать, что оптимальная величина параметра λ действительно существенно зависит от характеристик речи: относительные значения λ для читаемого материала настроечной выборки в [20] в среднем на треть меньше, чем для спонтанной речи из [21]. Более того, для обучающих данных среднее значение параметра λ оказывается ненулевым (0,12), как можно было бы ожидать. Для одного и того же типа материала наиболее существенное увеличение λ наблюдается для признаков окончания слов и функциональных слов (предлогах). Использование интерполированных моделей даже с усредненными (без классификации деревом решений) значениями λ приводит к повышению оценок правдоподобия данных для корректных гипотез, поэтому можно ожидать также снижения уровня ошибок распознавания при использовании сглаженных акустических моделей.

4 Заключение

Статья посвящена исследованию возможности снижения уровня ошибок при автоматическом распознавании естественной речи за счет использования неявных моделей вариативности произношения. В качестве основного источника вариативности рассматривается нечеткая, неполная артикуляция как следствие нарушения синхронизации работы частей речеобразующего тракта. Для учета такого типа вариативности предложено заменить исходные акустические модели их комбинациями в виде сглаживания параметров текущих моделей параметрами последующих. В ходе предварительных численных экспериментов на корпусах данных показано, что вариативность произнесения корректно рассматривать как временный фактор, обусловленный текущим фонетическим, позиционным и просодическим контекстом, соответственно параметры сглаживания также должны быть контекстно-зависимыми. Приведены контекстные признаки появления вариативности и показано, что использование сглаженных моделей звуков в потенциально вариативных позициях повышает правдоподобие данных, что коррелирует со снижением уровня ошибок при распознавании разговорной речи.

Литература

[1] Word error rate. `http://en.wikipedia.org/wiki/Word_error_rate`.

[2] *McAllaster D., Gillick L., Scattone F., Newman M.* Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch // Conference (International) on Speech and Language Processing. — Sydney, 1998. P. 1847–1850.

[3] *Saraclar M., Nock H., Khudanpur S.* Pronunciation modeling by sharing Gaussian densities across phonetic models // Comput. Speech Lang., 2000. Vol. 14. No. 4. P. 137–160.

[4] *Wester M.* Pronunciation modeling for ASR — knowledge-based and data-derived methods // Comput. Speech Lang., 2003. Vol. 17. P. 69–85.

[5] *Saraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // Comput. Speech Lang., 2004. Vol. 18(4). P. 375–395.

[6] *Jelinek F.* Statistical methods for speech recognition. — Cambridge, MA, USA: MIT Press, 1997. 305 p.

[7] *Lehr M., Gorman K., Shafran I.* Discriminative pronunciation modeling for dialectal speech recognition // International Speech Communication Association, Interspeech Conference Proceedings. Singapoure, 2014. P. 1458–1462.

[8] *Byrne B., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C., Zavaliagkos G.* Pronunciation modelling for conversational speech recognition: A status report from WS97 // IEEE Workshop on Automatic Speech Recognition and Understanding. — USA, 1997. P. 26–33. doi: 10.1109/ASRU.1997.659004.

[9] *Hitchinson B., Droppo J.* Learning non-parametric models of pronunciation in automatic speech recognition // Conference (International) on Acoustics, Speech, and Signal Processing Proceedings. — USA, 2011. P. 4904–4907.

[10] *Schramm H.* Modeling spontaneous speech variability for large vocabulary continuous speech recognition. Germany: Technical University of Aachen, 2006. D.Sc. Diss.

[11] *Livescu L., Glass J.* Feature-based pronunciation modeling for speech recognition // Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics Proceedings. — New York, NY, USA, 2004.

[12] *Hain T, Woodland P. C.* Dynamic HMM selection for continuous speech recognition // Proc. EuroSpeech, 1999. P. 1327–1330.

[13] *Hain T.* Implicit modelling of pronunciation variation in automatic speech recognition // Speech Commun., 2005. Vol. 46. P. 171–188.

[14] *Zheng J., Franco H., Stolcke A.* Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // Speech Communication, 2003. Vol. 41. P. 273–285.

[15] *Liu Y.* Modeling partial pronunciation variations for spontaneous Mandarin speech recognition // Comput. Speech Lang., 2003. Vol. 17. No. 4. P. 357–379.

[16] *Spiess T., Wrede B., Fink G. A., Kummert F.* Data-driven pronunciation modeling for ASR using acoustic subword units // Conference (International) on InterSpeech, 2003. P. 2549–2552.

[17] *Ostendorf M., Shafran I., Bates R.* Prosody models for conversational speech recognition // 2nd Plenary Meeting and Symposium on Prosody and Speech Processing. — USA, 2003. P. 147–154.

[18] *Rabiner L., Biing-Hwang J.* Fundamentals of speech recognition. — Signal processing ser. — New Jersey, USA: Prentice Hall, 1993. 496 p.

[19] *Чучупал В. Я., Маковкин К. А, Чичагов А. В., Кузнецов В. Б., Огарышев В. Ф.* Речевой корпус данных TeCoRus. Свидетельство об официальной регистрации базы данных № 2005620205, 2005.

[20] *Bogdanov D. S., Krivnova O. F., Podrabinovitch A. J., Arlazarov V. L.* Creation of Russian Speech Databases: Design, processing, development tools // Conference (International) on Speech and Computers Proceedings. Moscow, 2004. С. 650–656.

[21] База фрагментов разговорной русской речи. Свидетельство о регистрации базы данных 2016620687, 2016.

Поступила в редакцию 01.09.2016

Implicit pronunciation variation model for automatic speech recognition*

V. J. Chuchupal
v.chuchupal@gmail.com
Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

The variations in pronunciation of words in natural speech are considered as one of the main sources of speech recognition errors. This is the reason for development and implementation of the advanced pronunciation models in modern ASR (automatic speech recognition) systems. The paper considers the pronunciation variations that are caused by a fuzzy or an incomplete articulation that is frequently observed in spontaneous speech. The author proposes the use of the implicit pronunciation model that is implemented as the combination of the acoustical models of the adjacent phones. Such a model could be realized by smoothing or interpolation of the corresponding model parameters. Also, it is proposed to use the context-dependent interpolation, so that the values of the smoothing parameters are conditioned by the current position, syntax, and prosodic contexts of the sound. While the pronunciation modeling approach on the base of combination of acoustical models (including the interpolation) has already been discussed in literature, the proposed method based on the combination of the adjacent models with the use of the context-dependent smoothing parameters has not already been published as far as the author knows. The numerical experiments on the databases that contained both the read and spontaneous speech showed the correctness of the proposal about the use of the acoustic model combination on the base of interpolation and proposal to utilize the variable smoothing parameters such that the parameter values are conditioned on the features of phonemic context, syntax, and prosody.

Keywords: *automatic speech recognition; pronunciation variation; pronunciation modeling; hidden markov models*

DOI: 10.21469/22233792.2.4.01

References

[1] Word error rate. Available at: `http://en.wikipedia.org/wiki/Word_error_rate` (accessed January 10, 2017).

[2] McAllaster, D., L. Gillick, F. Scattone, and M. Newman. 1988. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. *Conference (International) on Speech and Language Processing*. Sydney. 1847–1850.

[3] Saraclar, M., H. Nock, and S. Khudanpur. 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. *Comput. Speech Lang.* 14(4):137–160.

[4] Wester, M. 2003. Pronunciation modeling for ASR — knowledge-based and data-derived methods. *Comput. Speech Lang.* 17:69–85.

[5] Saraclar, M., and S. Khudanpur. 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Comput. Speech Lang.* 18(4):375–395.

[6] Jelinek, F. *Statistical methods for speech recognition.* Cambridge, MA: MIT Press, 1997. 305 p.

[7] Lehr, M., K. Gorman, and I. Shafran. 2014. Discriminative pronunciation modeling for dialectal speech recognition. *International Speech Communication Association, Interspeech Conference Proceedings*. Singapoure. 1458–1462.

*The research was supported by the Russian Foundation for Basic Research (grant 14-01-00607).

[8] Byrne, B., M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos. 1997. Pronunciation modelling for conversational speech recognition: A status report from WS97. *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 26–33. doi: 10.1109/ASRU.1997.659004.

[9] Hitchinson, B., and J. Droppo. 2011. Learning non-parametric models of pronunciation in automatic speech recognition. *Conference (International) on Acoustics, Speech, and Signal Processing Proceedings*. USA. 4904–4907.

[10] Schramm, H. Modeling spontaneous speech variability for large vocabulary continuous speech recognition. Gernany: Technical University of Aachen. D.Sc. Diss.

[11] Livescu, L., and J. Glass. 2004. Feature-based pronunciation modeling for speech recognition. *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics Proceedings*. New York, NY.

[12] Hain, T., and P. C. Woodland. 1999. Dynamic HMM selection for continuous speech recognition. *Proc. EuroSpeech*. 1327–1330.

[13] Hain, T. 2005. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Commun.* 46:171–188.

[14] Zheng, J., H. Franco, and A. Stolcke. 2003. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition. *Speech Commun.* 41:273–285.

[15] Liu, Y. 2003. Modeling partial pronunciation variations for spontaneous Mandarin speech recognition. *Comput. Speech Lang.* 17(4):357–379.

[16] Spiess, T., B. Wrede, G. A. Fink, and F. Kummert. 2003. Data-driven pronunciation modeling for ASR using acoustic subword units. *Conference (International) on InterSpeech*. 2549–2552.

[17] Ostendorf, M., I. Shafran, and R. Bates. 2003. Prosody models for conversational speech recognition. *2nd Plenary Meeting and Symposium on Prosody and Speech Processing*. USA. 147–154.

[18] Rabiner, L., and J. Biing-Hwang. 1993. *Fundamentals of speech recognition*. Signal processing ser. New Jersey: Prentice Hall. 496 p.

[19] Chuchupal, V. J., K. A. Makovkin, A. V. Chichagov, V. B. Kuznetsov, and V. F. Ogaryshev. 2005. Speech corpus TeCoRus. The sertificate of database registration No. 2005620205. RosPatent.

[20] Bogdanov, D. S., O. F. Krivnova, A. J. Podrabinovitch, and V. L. Arlazarov. 2004. Creation of Russian Speech Databases: Design, processing, development tools. *Conference (International) on Speech and Computers Proceedings*. Moscow. 650–656.

[21] Corpus of spontaneous speech in Russian. 2016. The sertificate of database registration No. 2016620687. RosPatent.

Received September 1, 2016

Реализация пакета программ для встраивания и извлечения скрытых сообщений в аудиофайлах

А. А. Жарких[1]***, А. В. Горбунов***[2]
zharkikh090107@mail.ru; lergex@gmail.com
[1]Мурманский государственный технический университет
Россия, г. Мурманск, ул. Спортивная, 13
[2]МФ ФГБУ «Центр системы мониторинга рыболовства и связи»
Россия, г. Мурманск, ул. Траловая, 43

Представлены результаты разработки пакета программ для встраивания, извлечения, обнаружения и прочтения сообщений в аудиофайлах. Методология работы и программная реализация относятся к стеганографии — одному из основных направлений защиты информации. Стегоконтейнер представляет собой аудиофайл, получаемый в результате модификации контейнера последовательностью бит сообщения. Контейнер и стегоконтейнер представляют собой последовательности отсчетов импульсно-кодовой модуляции (ИКМ), а встраивание осуществляется простым суммированием отсчетов контейнера с отсчетами сигнала сообщения. Сигнал сообщения модулируется методом двоичной дискретной частотной манипуляции (ЧМ; английский термин — frequency shift keying, FSK), а контейнер перед встраиванием подвергается режекторной фильтрации. И методология, и пакет программ являются основой построения стеганографических систем со скачками по частоте (СЧ) и с вариацией просодических параметров речи.

Ключевые слова: *защита информации; стеганография; аудиосигналы; обнаружение сигналов; различение сигналов*

DOI: 10.21469/22233792.2.4.02

1 Введение

Цель данной работы — описание стеганографического метода, алгоритмов его реализации и программных средств для встраивания и извлечения. Метод базируется на модификации вектора отсчетов аудиосигнала, представленного в формате ИКМ. Сообщение представляет собой вектор отсчетов сигнала, полученного путем двоичной ЧМ из вектора бит сообщения.

Стеганография отличается от криптографии. Криптография изучает методы защиты информации, при которых допускается модификация сообщений (шифрование), стеганография же изучает методы защиты информации, в которых сообщение не изменяется (не шифруется), зато скрывается сам факт ее наличия. В стеганографии можно отметить три следующих основных раздела: методы встраивания сообщений в контейнеры; методы извлечения сообщений из контейнера; стегоанализ.

Наименее проработанная область исследования стеганографии — стегоанализ. Стегоанализ — это раздел стеганографии, который представляет собой совокупность методов обнаружения, различения и прочтения сообщений в различных контейнерах любой природы [1–7].

Необходимость обнаружения скрытых вложений может быть вызвана различными причинами. Данные средств массовой информации указывают, что зачастую стегосистемы разрабатываются организованной преступностью и террористами для решения своих незаконных задач (создание скрытых каналов связи, каналов утечки конфиденциальной информации) [**?**, 9]. В этом случае стегоанализ таких систем решает задачи, позитивные для

общества. Стегоанализ проводится с целью обнаружения, различения, прочтения и определения скрытых сообщений для предотвращения противоправной деятельности. Также одним из направлений применения стегоанализа является стегоанализ разработчиков стегосистем. Важным показателем стегосистемы является устойчивость при различных атаках нелегальных пользователей, а также воздействия дефектов носителей и помех в каналах передачи. Данное направление используется разработчиками легальных стегосистем для оценки их устойчивости к обнаружению, различению, прочтению и определению смысла скрываемого сообщения.

В настоящее время в направлении науки, посвященном методам стеганографии работают многие отечественные и зарубежные ученые: В. Г. Грибунин, И. Н. Оков, Б. Я. Рябко, И. В. Туринцев, А. В. Аграновский, А. Н. Фионов, В. Бендер (W. Bender), Н. Моримото (N. Morimoto) и др. В данной работе авторы описывают опыт разработки пакета программ, реализующего алгоритмы встраивания, извлечения и обнаружения сообщений в аудиофайлах. Работа сдержит три основных раздела.

В первом разделе формулируются требования к программному средству, реализующему извлечение и обнаружение произвольной структуры в аудиофайлах. Обоснован формат сообщения и аудиофайла контейнера.

Во втором разделе описана структура аудиофайла контейнера, скрываемого сообщения, а также методы встраивания извлечения и обнаружения, лежащие в основе программного средства.

В третьем разделе описана архитектура программных средств и их функциональные возможности.

В заключении отмечены перспективы развития и использования как рассмотренных методов, так и созданного пакета программ.

2 Постановка задачи и требования к программному средству

Необходимо выбрать метод встраивания, и для выбранного метода встраивания должен быть реализован метод обнаружения. Для этого нужно разработать с помощью пакета MATLAB функцию, реализующую встраивание скрытых сообщений в аудиофайлы формата wav. Также должно быть разработано программное средство на языке высокого уровня, предназначенное для обнаружения скрытых вложений различного типа в аудиофайлы различной структуры.

Пользователю предлагается некоторый аудиофайл формата wav. Необходимо определить, содержится ли в аудиофайле некоторое скрытое сообщение. Обнаружение ведется «полуслепым» методом, т. е. предполагается, что известна некоторая дополнительная информация о стегосистеме:

1) метод встраивания;
2) параметры метода встраивания (все или несколько).

При всех известных ключевых параметрах программа должна позволить пользователю прочесть скрытое сообщение. Если содержание скрытого сообщения известно пользователю, программа должна, при заданных параметрах обнаружения, определять долю правильно обнаруженных и необнаруженных бит, а также неправильно обнаруженных и необнаруженных бит сообщения.

2.1 Требования к структуре и функционированию

Структура программного средства должна быть модульной. Функциональная структура средства должна включать следующие модули:

– модуль извлечения;
– модуль обнаружения;
– модуль воспроизведения. Позволяет проигрывать файл-контейнер как аудиофайл. Используется для субъективной оценки качества встраивания;
– модуль справки. Выводит справочную информацию;
– модуль обработки ошибок. Обрабатывает исключения.

Программное средство должно допускать наращивание функциональных возможностей. Пользовательский интерфейс должен быть графическим и обеспечивать удобную и простую навигацию в диалоге с пользователем. Необходимо наличие справки (помощи). Программа должна функционировать под операционной системой Windows Vista/7/8/10, так как данная операционная система является наиболее распространенной среди пользователей персональных компьютеров. Необходим установленный в системе Microsoft .NET Framework 4.5.

2.2 Требования к оборудованию

К оборудованию предъявляются следующие требования:

– процессор с тактовой частотой 1 ГГц или выше;
– опретивное запоминающее устройство объемом 512 МБ;
– 50 МБ доступного пространства на жестком диске;
– Наличие звуковой карты (необязательно).

Основным языком взаимодействия пользователей и системы является русский язык. Графический интерфейс пользователя должен быть создан на русском языке.

3 Метод встраивания сообщений в ИКМ — отсчеты аудиосигнала с использованием двухпозиционной частотной манипуляции

Цифровая запись аудиосигналов базируется на выполнении двух операций над аналоговыми сигналами: дискретизации и квантования. В реальных устройствах эти операции осуществляются одновременно. В результате дискретизации аналоговый аудиосигнал заменяется последовательностью отсчетов, а в результате квантования каждый отсчет заменяется последовательностью бит. В итоге исходный аналоговый сигнал заменяется массивом целых чисел, каждое из которых ограничено числом разрядов, равным числу бит квантования.

Для восстановления аналоговой формы сигнала кроме данного массива необходимо знать число уровней квантования и интервал дискретизации.

Часто цифровое представление аудиосигнала называют импульсно-кодовой модуляцией аудиосигнала [10–12]. Это связано с тем, что грубую копию аудиосигнала можно представить как последовательность прямоугольных импульсов одинаковой длительности равной интервалу дискретизации, при этом амплитуда текущего импульса равна текущему значению элемента массива цифровой записи.

Авторы предлагают использовать массив отсчетов аудиосигналов как контейнер для встраивания сообщений.

В следствие разнообразия приложений, скрываемые сообщения могут иметь различную природу и структуру, поэтому договоримся далее, что любое сообщение перед вложением в контейнер преобразуется просто в последовательность бит. Единственное ограничение, предъявляемое к данной последовательности, — это число ее элементов, которое ограничивается емкостью контейнера и для различных методов встраивания может оказаться различным.

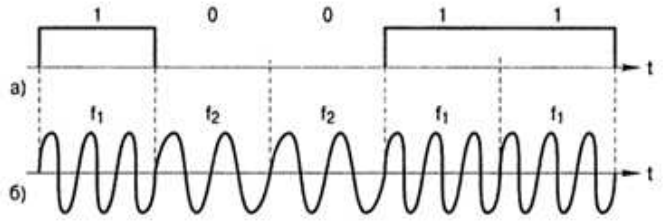


Рис. 1 Форма сигнала при ЧМ: (*а*) манипулирующий сигнал; (*б*) частотно-манипулирующий сигнал — радиосигнал ЧМ

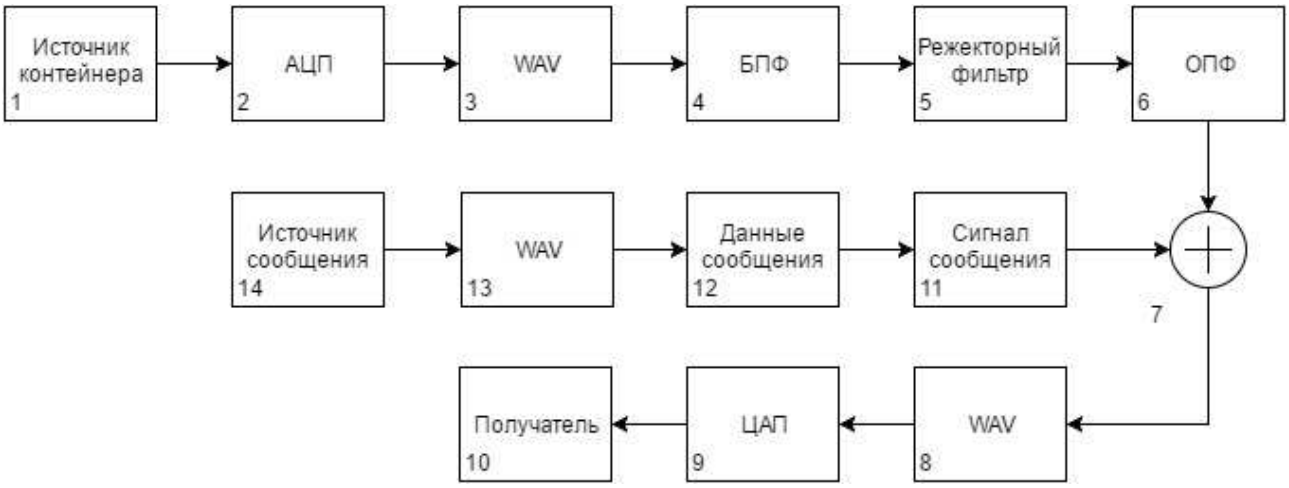


Рис. 2 Структурная схема алгоритма встраивания

При ЧМ каждому возможному значению передаваемого символа сопоставляется своя частота (рис. 1). В течение каждого символьного интервала передается гармоническое колебание с частотой, соответствующей текущему символу [13].

Алгоритм метода встраивания можно представить 14 блоками (рис. 2).

Структурная схема содержит следующие блоки:

1) источник контейнера;
2) аналого-цифровой преобразователь (АЦП);
3) формирователь wav файла контейнера;
4) блок быстрого преобразования Фурье (БПФ);
5) режекторный фильтр;
6) блок обратного преобразования Фурье (ОПФ);
7) сумматор;
8) формирователь wav файла стегоконтейнера;
9) цифроаналоговый преобразователь (ЦАП);
10) получатель данных;
11) сигнал сообщения;
12) данные сообщения;
13) отсчеты сообщения;
14) источник сообщения.

Блок 1 — источник контейнера, т. е. аудиосигнал, имеющий аналоговую форму. Из блока 1 сигнал поступает в блок 2, который представляет собой АЦП. На выходе бло-

ка 2 получается цифровой сигнал. Существует блок 3, который записывает выход из блока 2 в wav файл без сжатия. Из блока 3 данные поступают в блок 4, где осуществляется БПФ. Результат преобразования Фурье передается на блок 5, который представляет собой режекторный фильтр, вырезающий часть спектральной составляющей из спектра сигнала-контейнера. Результат преобразования контейнера подается на устройство 6 — ОПФ. Далее в сумматоре 7 осуществляется сложение контейнера и сообщения.

Формирование сигнала сообщения начинается в блоке 14. Блок 14 представляет собой источник битов сообщения. Источник битов сообщения преобразуется в некоторый набор отсчетов сигнала. Этот набор отсчетов объединяется в блоке 12 с данными о контейнере, полученными в блоке 5. Объединенные данные из блока 12 используются для формирования сигнала сообщения в блоке 11. Сигнал сообщения поступает на сумматор 7, где складывается с контейнером, поступающим с блока 5, как было сказано ранее. Результат суммирования подается в блок 8, где формируется файл стегоконтейнера с расширением wav. При необходимости воспроизведения стегоконтейнера в целом файл с устройства 8 преобразуется в ЦАП 9 и далее воспринимается получателем 10.

Сам алгоритм встраивания можно описать следующим образом. Аудиофайл-контейнер разбивается на отрезки, содержащие число отчетов, кратное степени двойки. Производится преобразование Фурье отрезка. Для встраивания данных используются узкие полосы частот вблизи выбранных частот. Для этих узких полос оценивается энергия, исходя из которой выбирается амплитуда модулированного скрываемого сообщения.

Модулированное сообщение формируется следующим образом:

1) выбираются две частоты, лежащие внутри полос встраивания;
2) оценивается длительность встраивания одного бита сообщения — в это время должно уложиться несколько периодов частоты встраивания;
3) формируется добавочное сообщение: встраиваемой единице соответствует колебание нижней частоты из полосы встраивания, а нулю — верхней частоты из полосы встраивания;
4) амплитуда модулированного сообщения выбирается такой, чтобы энергия встраиваемого сообщения соответствовала энергии контейнера в полосе встраивания.

После формирования сообщения полосы встраивания режектируются из спектра исходного отрезка и производится ОПФ. Далее полученные отсчеты суммируются с отсчетами модулированного сообщения. Ширина полос встраивания, а также частоты встраивания оцениваются исходя из требуемой скорости передачи бит скрываемого сообщения, а также из соображений незаметности встраивания.

На рис. 3 представлена структурная схема алгоритма метода, в котором происходит извлечения сообщения.

Структурная схема содержит следующие блоки:

1) источник стего;
2) АЦП;
3) формирователь wav файла стегоконтейнера;
4) блок БПФ;
5) полосовой фильтр;
6) блок ОПФ;
7) блок формирования сигнала сообщения;
8) данные о сообщении;
9) отсчеты файла сообщения;

Рис. 3 Структурная схема алгоритма извлечения

10) получатель отсчетов сообщения.

Для извлечения скрытого сообщения необходимо знать длину отрезка встраивания, частоту встраивания, а также ширину полос встраивания. Аудиофайл-стегоконтейнер также разбивается на отрезки известной длины и производится БПФ. Анализируются полосы частот вблизи полос встраивания. Полосы встраивания отфильтровываются, и производится ОПФ полученного узкополосного сигнала. Таким образом получаем модулированное сообщение. Далее оценивается длительность передачи одного бита сообщения. Модулированное сообщение разбивается на отрезки передачи одного бита, внутри каждого из которых оценивается частота и по этой частоте принимается решение о переданном бите. Таким образом принимается скрываемое сообщение.

Прием частотно-манипулированного сигнала осуществляется корреляционным методом. Корреляционный прием может быть когерентным или некогерентным. В данной работе используется когерентный прием. Он используется, если известны начальные фазы посылок. Его принцип состоит в вычислении взаимной корреляции между принимаемым сигналом и колебаниями-образцами (опорными сигналами), представляющими собой гармонические колебания с используемыми для манипуляции частотами.

Взаимная корреляция сигнала с k-м опорным сигналом для n-го по времени символа рассчитывается следующим образом:

$$u_k(n) = \int\limits_{nT}^{(n+1)T} s(t)\cos(\omega_k t + \varphi_{0k})\,dx,$$

где $s(t)$ — частотно-манипулированный сигнала; ω_k — частота манипуляции, соответствующая символу, равному k; φ_{0k} — начальная фаза посылки; T — длительность передачи символа.

Использованные пределы интегрирования задают обработку n-го символа. При программной реализации демодуляции частотно-манипулированного сигнала вместо интегрирования необходимо использовать суммирование дискретных отсчетов подынтегрального выражения [**?**].

Схема обнаружителя приведена на рис. 4. Сигнал сообщения $s(t)$ перемножается с опорными сигналами для нулевого $\cos(\omega_0 t + \varphi_0)$ и единичного $\cos(\omega_1 t + \varphi_1)$ битов. Затем результаты интегрируются (суммируются) и поступают в блок сравнения. В этом бло-

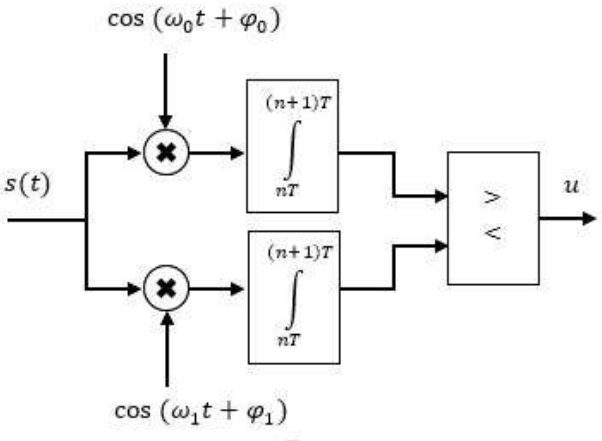


Рис. 4 Структурная схема обнаружения

ке полученные суммы сравниваются с пороговым значением. Порог выбирается с учетом энергии опорного сигнала и составляет его половину.

Энергия сигнала для нулевого бита:

$$E_{s_0} = \sum_{k=0}^{T} S_{0_k}^2 \Delta t = \frac{A_0^2 T}{2},$$

где S_{0_k} — значение сигнала нулевого бита в k-м отсчете; T — длительность одного бита; Δt — приращение отсчетов; A_0 — амплитуда сигнала.

Энергия сигнала для бита, равного единице:

$$E_{s_1} = \sum_{k=0}^{T} S_{1_k}^2 \Delta t = \frac{A_1^2 T}{2},$$

где S_{1_k} — значение сигнала бита равного единице в k-м отсчете; A_1 — амплитуда сигнала.

Пороговые значения, таким образом составляют $E_{s_0}/2$ и $E_{s_1}/2$ для бита «0» и бита «1» соответственно. Решение об обнаруженном бите в блоке сравнения принимается исходя из следующего алгоритма:

1) если сумма для бита 0 больше своего порогового значения, а сумма бита 1 меньше своего, то считается, что встроен бит 0;
2) если сумма для бита 1 больше своего порогового значения, а сумма бита 0 меньше своего, то считается, что встроен бит 1;
3) если суммы для обоих битов меньше их порогов, то делается вывод, что встраивания не было;
4) если суммы для обоих битов больше своих пороговых значений, то выбирается бит, сумма которого больше.

Таким образом на выходе обнаружителя формируется цепочка из нулей, единиц и символа «N» — отсутствия встраивания.

4 Архитектура программного средства

На основании изложенных в разд. 3 алгоритмов были разработаны функция в пакете MATLAB, осуществляющая встраивание, а также программное средство на языке C#,

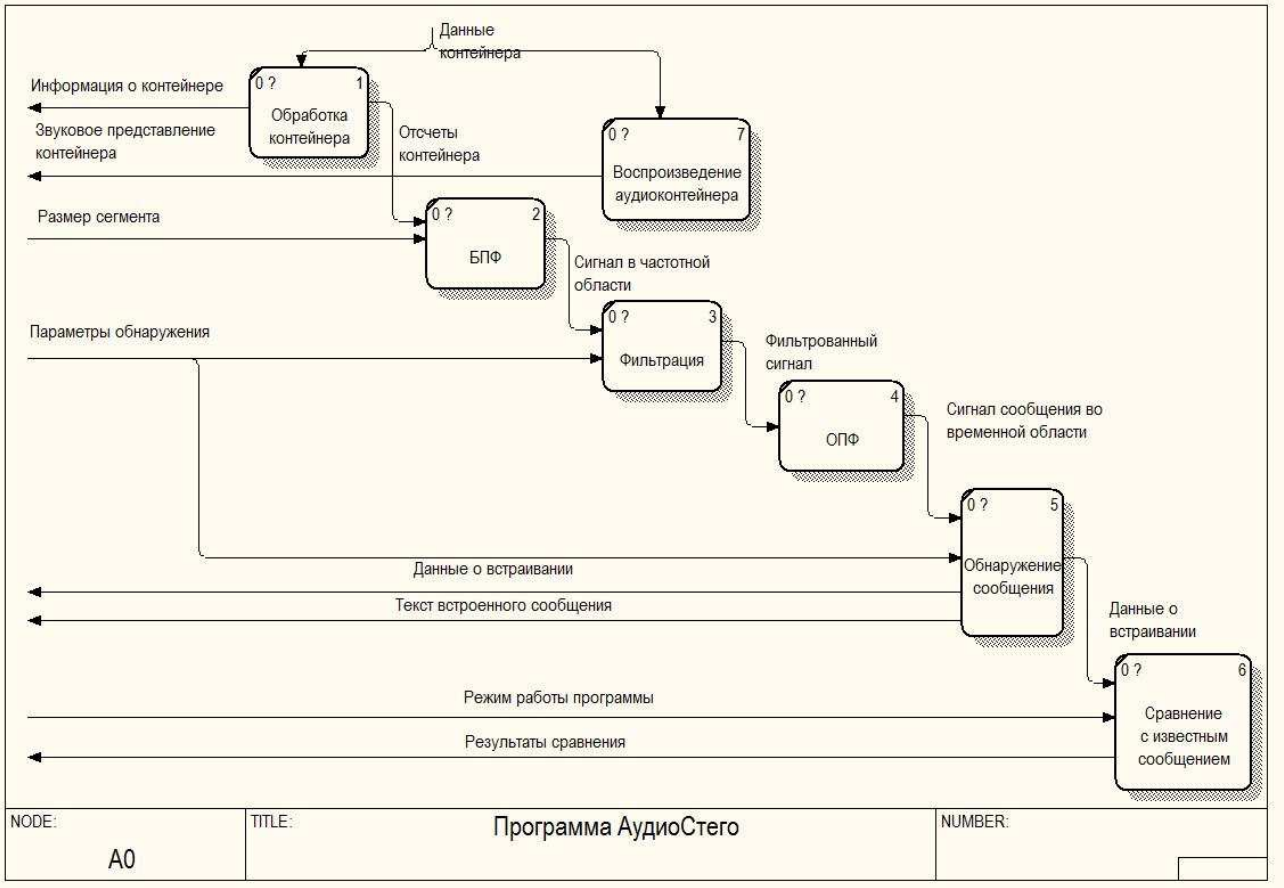


Рис. 5 Контекстная диаграмма потоков данных программы обнаружения

позволяющее обнаруживать скрытые методом на основе двухпозиционной ЧМ вложения в аудиофайле.

Функция встраивания позволяет осуществлять скрытие в контейнере произвольного (ограниченного емкостью контейнера) массива битов сообщения. Для этого задается аудиофайл-контейнер и выбираются параметры встраивания, такие как длина сегмента, частоты нулевого и единичного бита, длительность одного бита. На выходе получается стегоконтейнер в формате wav, содержащий сообщение.

Структурная схема программы для обнаружения в виде контекстной диаграммы потоков данных приведена на рис. 5.

Эта диаграмма содержит 7 процессов:

1) обработка контейнера — включает чтение с диска аудиофайла-контейнера, чтение и интерпретацию его заголовочной информации, определение формата хранения отсчетов сигнала, преобразование их в числа с плавающей точкой двойной точности;
2) БПФ — производит БПФ для последующей фильтрации сигнала в частотной области;
3) фильтрация — применяет к результату предыдущего процесса полосно-пропускающий фильтр, выделяя тем самым предполагаемый сигнал скрытого сообщения;
4) ОПФ — переводит сигнал сообщения из предыдущего процесса во временную область;
5) обнаружение сообщения — производит обнаружение битов скрытого сообщения в соответствии с заданными пользователем параметрами. Возвращает информацию о встраивании и, если возможно, текстовое представление полученного сообщения;

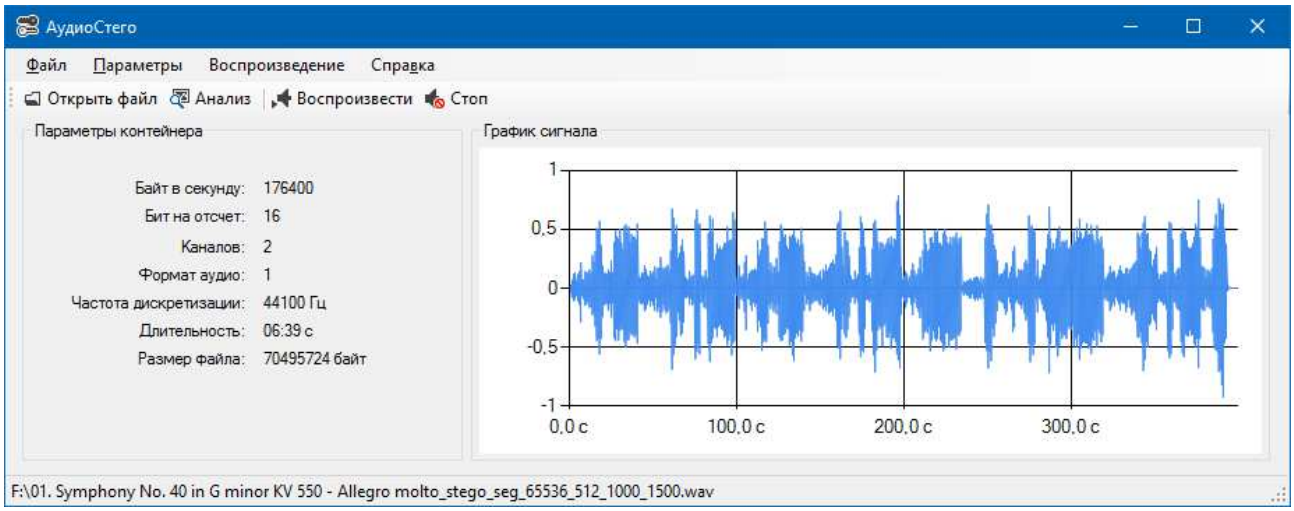


Рис. 6 Главное окно программы

6) сравнение с известным сообщением — сравнивает полученное сообщение с известным и выводит результаты оценки правильности обнаружения;
7) воспроизведение аудиоконтейнера — проигрывает аудиофайл.

К входной информации для программного средства относятся различные параметры, определяемые пользователем, а также аудиофайл — потенциальный стегоконтейнер. Параметры обнаружения включают в себя значения:

1) амплитуды;
2) частоты нуля;
3) частоты единицы;
4) длительности 1 бита;
5) порога обнаружения.

К выходной информации относятся информация о контейнере, результаты работы алгоритма обнаружения, текстовое представление обнаруженного сообщения. Для тестового режима работы программы дополнительно выводятся результаты сравнения с известным сообщением. Имеется возможность воспроизвести и прослушать открытый аудиофайл.

Текст встроенного сообщения представляет собой строковое представление полученного скрытого сообщения в кодировке windows-1251. Данные о встраивании состоят из информации по общему количеству бит, количеству встроенных бит, количеству бит без встраивания, а также числу нулевых и единичных битов.

Результаты сравнения содержат сравнительные оценки результатов обнаружения. К этим оценкам относятся количество правильно необнаруженных бит, правильно обнаруженных и различенных бит, правильно обнаруженных, но неправильно различенных бит, а также количество ложных обнаружений и ложных необнаружений.

Главное окно программы изображено на рис. 6. Здесь отображается основная информация о контейнере: количество каналов, частота дискретизации, продолжительность звучания, размер и другие параметры, а также графическое представление сигнала выбранного канала.

Окно «Стегоанализ» используется для анализа открытого контейнера (рис. 7).

В этом окне пользователем задаются основные параметры для обнаружения вложения. После выполнения процедуры обнаружения выводится статистика по общему количеству

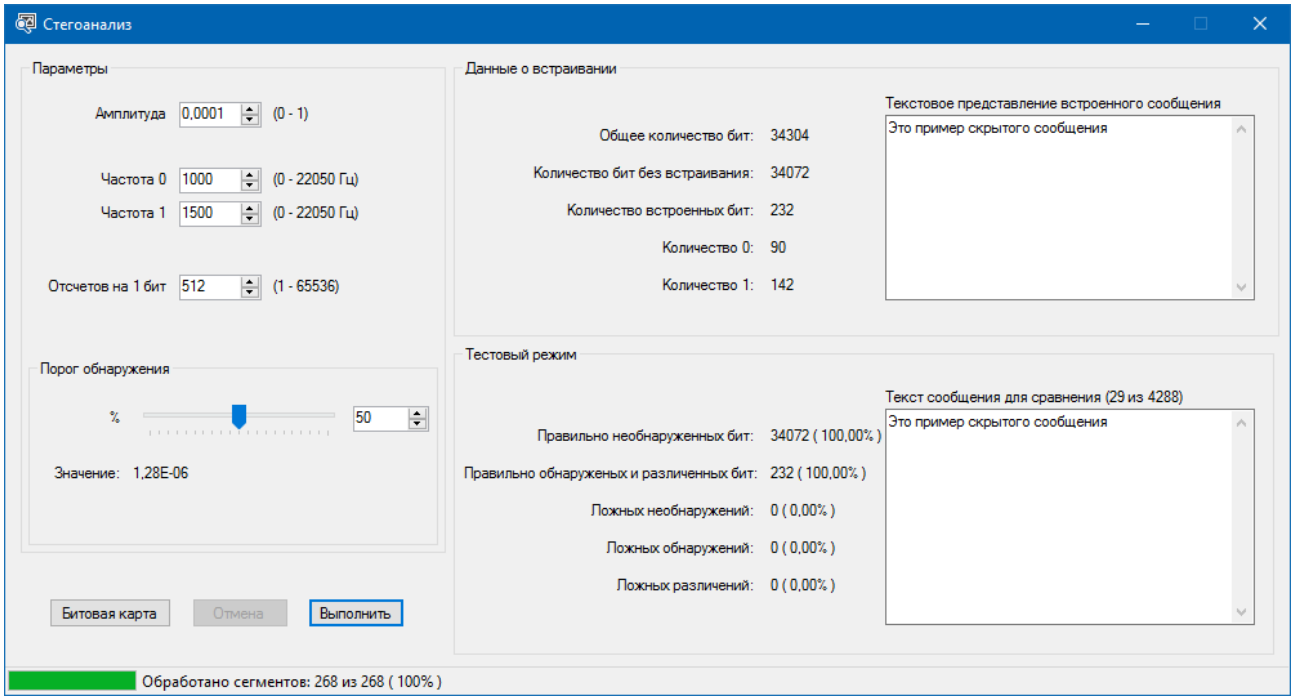


Рис. 7 Окно «Стегоанализ»

бит, количеству встроенных бит, количеству бит без встраивания, а также числу нулевых и единичных битов. В текстовое поле выводится строковое представление обнаруженного встроенного сообщения. При известном содержании скрытого сообщения после исполнения программа отобразит сравнительные оценки результатов обнаружения. К таким оценкам относятся количество правильно необнаруженных бит, правильно обнаруженных и различенных бит, правильно обнаруженных, но неправильно различенных бит, а также количество ложных обнаружений и ложных необнаружений.

5 Заключение

Завершая вышесказанное, отметим основные результаты работы.

Был создан пакет программ, который в широком диапазоне параметров аудиофайла и различных вариантов сообщений позволяет осуществлять встраивание, извлечение, обнаружение и прочтение сообщений в аудиофайлах.

Данный пакет по сути является виртуальной экспериментальной установкой для исследования различных характеристик стеганографических систем, использующих аудиофайлы-контейнеры. Он может быть использован по прямому назначению, т. е. для реализации всех четырех перечисленных функций, а также в учебном процессе по дисциплинам, связанным с защитой информации. Авторы полагают, что он также является весьма перспективным для научных направлений, таких как стеганография, интеллектуальный анализ данных и распознавание образов. По меньшей мере можно указать следующие перспективные области применения изложенных в работе методов и реализованного программного средства:

– методика стеганографии со скачками по частоте [14];
– методика стеганографии с модификацией просодических параметров речи [15];
– статистическая теория распознавания образов [16].

Литература

[1] *Bender W., Gruhl B., Morimoto N., Lu A.* Techniques for data hiding // IBM Syst. J., 1996. Vol. 35. № 3.

[2] *Чваркова И. Л.* Стеганографические методы скрытия информации в аудиоданных // Электроника, 2003. № 11. С. 54–56.

[3] *Романцов А. П., Бугаев В. С., Фролов М. А.* Комплекс лабораторных работ по стеганографии / Под ред. Заслуженного деятеля науки РФ д.т.н. проф. А. В. Петракова. — М.: РИО МТУСИ, 2005. 92 с.

[4] *Конахович Г. Ф., Пузыренко А. Ю.* Компьютерная стеганография. Теория и практика. — Киев: МК-Пресс, 2006. 288 с.

[5] *Аграновский А. В., Балакин А. В., Грибунин В. Г., Сапожников С. А.* Стеганография, цифровые водяные знаки и стеганоанализ. — М.: Вузовская книга, 2009. 220 с.

[6] *Грибунин В. Г., Оков И. Н., Туринцев И. В.* Цифровая стеганография. — М.: Солон-Пресс, 2009. 265 с.

[7] *Гурин А. В., Жарких А. А., Пластунов В. Ю.* Технологии встраивания цифровых водяных знаков в аудиосигнал / Под общ. ред. А. А. Жарких. — М.: Горячая линия — Телеком, 2015. 116 с.

[8] *Kelley J.* Terror groups hide behind Web encryption // USA Today, 2001. `http://usatoday30.usatoday.com/life/cyber/tech/2001-02-05-binladen.htm`.

[9] *Fox S.* FBI: Russian spies hid codes in online photos // NBC News, 2010. `http://www.nbcnews.com/id/38028696/ns/technolog_and_science-science/t/fbi-russian-spies-hidcodes-online-photos`.

[10] *Евсеев А. И., Сорокин П. М., Кончаловский В. Ю.* Преобразование непрерывных сигналов в дискретные // Передача информации. `http://peredacha-informacii.ru`.

[11] *Котельников В. А.* О пропускной способности эфира и проволоки в электросвязи — Всесоюзный энергетический комитет // Мат-лы к I Всесоюзному съезду по вопросам технической реконструкции дела связи и развития слаботочной промышленности, 1933. Репринт. УФН, 2006. Т. 176. № 7. С. 762–770.

[12] *Радзишевский А. Ю.* Основы аналогового и цифрового звука. — М.: Вильямс, 2006. 288 с.

[13] *Сергиенко А. В.* Цифровая обработка сигналов. — СПб.: Питер, 2002. 608 с.

[14] *Torrieri D.* Principles of spread-spectrum communication systems. — Boston, MA, USA: Springer Science, 2005. 456 p.

[15] *Потапова Р. К.* Речь: коммуникация, информация, кибернетика. — 4-е изд., доп. — М.: ЛИБРИКОМ, 2010. 594 с.

[16] *Фукунага К.* Введение в статистическую теорию распознавания образов / Пер. с англ. — М.: Наука, 1979. 368 с. (*Fukunaga K.* Introduction to statistical pattern recognition. — New York, NY, USA: Academic Press, 1972. 250 p.)

Поступила в редакцию 29.08.2016

Implementation of the software package for embedding and extracting hidden messages in audiofiles

A. A. Zharkikh[1] ***and A. V. Gorbunov***[2]
zharkikh090107@mail.ru; lergex@gmail.com
[1]Murmansk State Technical University, 13 Sportivnaya Str., Murmansk, Russia
[2]Murmansk Branch SO CFMC, 43 Tralovaya Str., Murmansk, Russia

The results of development of a software for embedding, extraction, detection, and reading of messages in audiofiles are provided. The methodology of work and program implementation belong to a steganography — one of the main directions of information security. Stegocontainer represents the audiofile received as a result of modification of a container by the sequence of bits of the message. A container and a stegocontainer represent the sequences of counting of the pulse code modulation, and embedding is performed by simple summing of counting of a container with counting of a signal of the message. The signal of the message is modulated by the method of binary discrete frequency manipulation, and the container before embedding is exposed to rejection filtering. Both the methodology and the software are a basis for steganography systems with frequency hopping and with a variation of prosodic parameters of the speech creation.

Keywords: *information security; steganography; audio signals; signals detection; signals distinction*

DOI: 10.21469/22233792.2.4.02

References

[1] Bender, W., B. Gruhl, N. Morimoto, and A. Lu. 1996. Techniques for data hiding. *IBM Syst. J.* 35(3).

[2] Chvarkova, I. L. 2003. Steganograficheskie metody skrytiya informatsii v audiodannykh [Steganographic techniques to hide information in the audio data]. *Elektronika* [Electronics] 11:54–56.

[3] Romantsov, A. P., V. S. Bugaev, and M. A. Frolov. 2005. *Kompleks laboratornykh rabot po steganografii* [Complex laboratory works on steganography]. Moscow: RIO MTUSI. 92 p.

[4] Konakhovich, G. F., and A. Yu. Puzyrenko. 2006. *Komp'yuternaya steganografiya. Teoriya i praktika* [Computer steganography. Theory and practice]. Kiev: MK-Press. 288 p.

[5] Agranovskiy, A. V., A. V. Balakin, V. G. Gribunin, and S. A. Sapozhnikov. 2009. *Steganografiya, tsifrovye vodyanye znaki i steganoanaliz* [Steganography, digital watermarking, and steganalysis]. Moscow: Vuzovskaya kniga. 220 p.

[6] Gribunin, V. G., I. N. Okov, and I. V. Turintsev. 2009. *Tsifrovaya steganografiya* [Digital steganography]. Moscow: Solon-Press. 265 p.

[7] Gurin, A. V., A. A. Zharkikh, and V. Yu. Plastunov. 2015. *Tekhnologii vstraivaniya tsifrovykh vodyanykh znakov v audiosignal* [Technology of embedding digital watermarks into audiosignal]. Moscow: Goryachaya liniya — Telekom. 116 p.

[8] Kelley, J. 2001. Terror groups hide behind Web encryption. *USA Today*. Available at: `http://usatoday30.usatoday.com/life/cyber/tech/2001-02-05-binladen.htm` (accessed December 28, 2016).

[9] Fox, S. 2010. FBI: Russian spies hid codes in online photos. *NBC News*. Available at: `http://www.nbcnews.com/id/38028696/ns/technology_and_science-science/t/fbi-russian-spies-hidcodes-online-photos` (accessed December 28, 2016).

[10] Evseev, A. I., P. M. Sorokin, and V. Yu. Konchalovskiy. Preobrazovanie nepreryvnykh signalov v diskretnye [Convert continuous signals to discrete]. — Peredacha informatsii [Transmission of information]. Available at: `http://peredacha-informacii.ru` (accessed December 28, 2016).

[11] Kotel'nikov, V. A. 2006. O propusknoy sposobnosti efira i provoloki v elektrosvyazi [On the transmission capacity of "ether" and wire in electrocommunications]. *Mat-ly k I Vsesoyuznomu s"ezdu po voprosam tekhnicheskoy rekonstruktsii dela svyazi i razvitiya slabotochnoy promyshlennosti, 1933* [1st All-Union Conference on the Technological Reconstruction of the Communications Sector and the Development of Low-Current Engineering Proceedings]. Reprint. *UFN* 176(7):762–770.

[12] Radzishevskiy, A. Yu. *Osnovy analogovogo i tsifrovogo zvuka* [Foundations of analog and digital audio]. Moscow: Vil'yams. 288 p.

[13] Sergienko, A. V. 2002. *Tsifrovaya obrabotka signalov* [Digital signal processing]. SPb.: Piter. 608 p.

[14] Torrieri, D. 2005. *Principles of spread-spectrum communication systems*. Boston, MA: Springer Science. 456 p.

[15] Potapova, R. K. 2010. *Rech': Kommunikatsiya, informatsiya, kibernetika* [Speech: Communication, information, cybernetics]. Moscow: LIBRIKOM. 594 p.

[16] Fukunaga, K. 1972. *Introduction to statistical pattern recognition*. New York, NY: Academic Press. 250 p.

Received August 29, 2016

Сопоставление и интеграция подходов к дешифровке древнерусских знаменных песнопений*

И. В. Бахмутова, В. Д. Гусев, Л. А. Мирошниченко, Т. Н. Титкова
bakh@math.nsc.ru; gusev@math.nsc.ru; luba@math.nsc.ru; titkova@math.nsc.ru
Институт математики им. С. Л. Соболева Сибирского отделения РАН
Россия, г. Новосибирск, ул. акад. Коптюга, 4

Рассматриваются наиболее перспективные, по мнению авторов, подходы к проблеме нотолинейной реконструкции (дешифровки) древнерусских церковных песнопений XVII в. и выше, представленных в знаменной форме записи. Предпочтение отдается предложенному авторами подходу, основанному на использовании внутригласовых инвариантов (ВИ) — цепочек знамен, интерпретация которых характеризуется минимальным уровнем неоднозначности. Приводятся оценки эффективности разных подходов. Исследуется возможность их интеграции.

Ключевые слова: *двознаменники; гласы; знаменные песнопения; попевки; дешифровка; инварианты; квазиинварианты*

DOI: 10.21469/22233792.2.4.03

1 Введение

Проблема нотолинейной реконструкции (*дешифровки*) древнерусских церковных песнопений, представленных в знаменной форме записи, является одной из наиболее актуальных в музыкальной медиевистике [1]. «Читаемыми» (с определенными оговорками) считаются лишь тексты XVII в. и более позднего периода, в которых знамена снабжены специальными знаками — *пометами*. Различают степенные и указательные пометы. Первые уточняют высоту распева знамени, вторые — особенности его распева. Беспометные рукописи XVI в. и более раннего периода практически нечитаемы.

Общие сведения о *знаменном пении* можно найти в [2]. Это одноголосное (унисонное) пение, подчиняющееся системе осмогласия (см. разд. 2). Знамена интерпретируются цепочками нот разной длины (от одного до четырех–пяти нотных знаков, исключения редки). Начертания знамен не содержат в явном виде информацию о звуковысотной привязке их распева. Реконструкция именно этого параметра вызывает наибольшие затруднения. Лишь представители отдельных классов знамен, например «крюков», «статей», могут быть упорядочены по указанному параметру. Так, крюк простой () следует «возгласити мало повыше строки», крюк мрачный () — «паки повыше простого», крюк светлый () — «мрачного повыше», крюк тресветлый () — «светлого повыше» [3]. Даже такая информация носит относительный характер, поскольку в конкретном песнопении эти крюки могут входить в состав разных структурных единиц и определяющим фактором будет звуковысотная привязка самих структурных единиц, а не отдельных их компонентов.

Известные примеры дешифровки беспометной знаменной нотации весьма немногочисленны (см., например, [4–7]) и сделаны вручную. Они касаются отдельных песнопений (или узких классов песнопений), эволюцию которых можно проследить по архивным материалам в течение достаточно длительного периода. Существенную роль здесь играет наличие

*Работа выполнена при финансовой поддержке РФФИ, проект №16-07-00812 «Стратегия и начальная версия алгоритма дешифровки древнерусских знаменных песнопений».

графически близких византийских версий и позднерусской читаемой версии. Возникающий при этом вопрос о степени достоверности полученной реконструкции затрагивается (в лучшем случае) лишь на качественном уровне.

Сложность процесса дешифровки обусловлена многозначностью соответствия «знамя–нота». В зависимости от ряда факторов некоторые знамена могут иметь до 10 различных интерпретаций, отличающихся друг от друга интервально-ритмическими характеристиками. При этом каждая из интерпретаций допускает различные звуковысотные привязки.

При всей сложности процесса дешифровки некоторые его этапы могут быть автоматизированы. Укажем, в частности, на необходимость создания баз данных, содержащих пометные песнопения разных жанров, двознаменники (билингвы в формате «знамя–нота»), электронные азбуки и словари структурных единиц (попевки, лица, фиты). Требуется разрабатывать алгоритмы дешифровки с использованием электронных словарей, алгоритмы распознавания знаменной нотации, обнаружения ошибок, визуализации, редактирования и проигрывания мелодий. Некоторые шаги по реализации указанной программы действий описаны в [8].

В своем подходе к проблеме дешифровки авторы настоящей работы опираются на *двознаменники* (см. [1, гл. 12]). Это певческие рукописи конца XVII – начала XVIII вв., в которых песнопения записаны в виде трех синхронизированных параллельных текстов — знаменного, нотолинейного и стихотворного (старославянский язык). Часть из них используется для «обучения» (в данном случае для выбора информативной системы описания исходных данных в виде множества *внутригласовых инвариантов* (ВИ) и *квазиинвариантов* (КВИ) — см. разд. 2). Эти понятия введены авторами в [9] и использованы в [10] для создания начальной версии алгоритма дешифровки *беспометных* знаменных песнопений XVII – XVIII вв.

Другая часть двознаменников (с предварительно устраненными пометами) используется для «контроля». Это дает возможность количественно оценить эффективность подхода в виде доли знамен, интерпретированных «правильно», т. е. так, как это указано в двознаменнике. Такого рода показатели отсутствуют в цитируемых работах [4–7].

В данной статье основное внимание уделено сопоставлению различных подходов к дешифровке *беспометной* знаменной нотации. Более детально рассматриваются два подхода. Один из них, предлагаемый авторами, основан на использовании внутригласовых инварианов. Другой опирается на попевочную структуру песнопений, отраженную (с той или иной степенью полноты) в известной подборке В. М. Металлова [11]. Исследуется возможность интеграции обоих подходов. Описана текущая версия алгоритма, реализующая возможность интеграции указанных и ряда других подходов.

2 Основные понятия и обозначения

2.1 Система осмогласия

Знаменное пение регламентируется *системой осмогласия*. В древнегреческом церковном пении, лежащем в основе знаменного распева, этому понятию соответствовало пение на 8 ладов (гласов): дорийский, фригийский и т. д. Начало системе положил обычай в каждый из восьми дней Пасхи исполнять песнопения на особый лад. Восьмидневный цикл напевов, которые хор исполнял в унисон (монодия), был распространен затем на 8 недель (одноголосный напев конкретного дня повторялся в течение соответствующей ему по порядку недели). Восемь недель составляли «столп», который циклически повторялся в течение года.

Система осмогласия проявляет себя во многих отношениях. Одно и то же знамя может по-разному интерпретироваться в разных гласах. Существуют характерные только для конкретного гласа мелодические обороты (попевки). Попевочная техника формирования мелодий, сводящаяся к их *комбинированию* из регламентированной совокупности гласовых попевок, является ограничительной по своей сути и отличает знаменные песнопения от произведений, созданных композиторами. Гласы отличаются по «высотности», хотя формально этот параметр никто не определял. Отдельные пары гласов демонстрируют значимое сходство по многим параметрам. В частности, сходными считаются гласы 1 и 5, 2 и 6, 3 и 7, 4 и 8. Все эти моменты следует учитывать, а иногда и использовать при дешифровке.

2.2 Структурные единицы

Основными структурными единицами знаменного распева являются *попевки*. Это устойчиво повторяющиеся мелодические обороты, представленные цепочками из 3–7 знамен (в среднем). Обычно каждому знамени соответствует один слог распеваемого текста. Большая часть попевок *гласоспецифична*, т. е. характерна для какого-то конкретного гласа. Однако некоторые попевки встречаются в разных гласах. Вопросам структурной организации попевок посвящена работа [12]. Все попевки разделены по типам кадансов (завершений) на 24 группы, каждая из которых имеет свою первооснову (архетип) из трех (иногда двух) знамен. Архетипы допускают изменения в своем знаковом составе. Измененный архетип называется производным от основного. Архетипу предшествует цепочка из нескольких знамен (обычно не более четырех), называемая подводом. Знамена, завершающие попевку, по большей части интерпретируются целой нотой и относятся преимущественно к семейству статей ( , ,  ,  и др.).

Все множество гласовых попевок образуется комбинированием архетипа (и его производных) с вариантным подводом. Оценки суммарного (по всем гласам) числа попевок существенно расходятся. Подборка В. М. Металлова (одна из наиболее полных) содержит свыше 500 попевок, представленных в *нотолинейной форме* [11]. Оценка М. В. Бражникова, приведенная в [2], вдвое превышает металловскую (порядка 1000 попевок).

Кроме попевок в песнопениях встречаются весьма спрецифические («тайнозамкненные») мелодические обороты — так называемые «лица» и «фиты». Они служат в основном для украшения знаменного распева. Их изъятие из песнопения обедняет напев, но, в отличие от попевок, не разрушает его. Под «*тайнозамкненностью*» понимается «графический прием зашифровки напева посредством такого условного сочетания знамен, которое не образует этого напева в случае исполнения знамен в этом же последовании, но согласно их обычному певческому значению» [2]. Здесь напрашивается аналогия с идиоматическими выражениями в естественном языке, значение которых, как известно, также не определяется значением входящих в них слов, взятых по отдельности. Именно поэтому певческое значение лиц и фит обычно представляется «*разводом*», т. е. достаточно длинной (по сравнению с попевками) цепочкой «рядовых» знамен, трактуемых уже в обычном смысле. Существуют некоторые проблемы с определением границ лицевых и фитных начертаний в знаменном тексте, но эти вопросы выходят за рамки данной статьи. Здесь же важно отметить, что элемент «тайнозамкненности» до некоторой степени может сохраняться и в разводах лиц и фит, создавая определенные проблемы при дешифровке.

К перечисленным структурным единицам (попевки, лица, фиты) следует, по мнению авторов, добавить и *тандемные повторы* цепочек знамен длины $1, 2$ и т. д. вплоть до повторения попевок. Тандемы с малой длиной периода обычно встречаются на стыках

между попевками. Дешифровка тандемных повторов вызывает затруднения, поскольку повтору на знаменном уровне не всегда соответствует повтор на нотолинейном уровне.

2.3 Певческие книги

Певческие книги предназначены для фиксации в рукописной форме церковного пения, являющегося неотъемлемой частью богослужения. Типов певческих книг не слишком много. К основным относят Октоих, Ирмологий, Праздники, Минеи, Триодь. Часть книг содержит песнопения разных жанров (например, Октоих), другие являются моножанровыми (например, Ирмологий, Триодь). Однотипные певческие книги близки по содержанию, хотя точно совпадающих певческих книг, по-видимому, не существует.

В распоряжении авторов на данный момент имеются три двознаменных Октоиха конца XVII – начала XVIII вв. (Российская научная библиотека (РНБ), С.-Петербург, Соловецкое собрание, шифры 619/647, 618/644 и QI188), а также двознаменники «Праздники» (РНБ, С.-Петербург, Кирилло-Белозерское собрание, шифр 797/1054) и Ирмологий из собрания В. Ф. Одоевского (РГБ, Москва, Ф.210, №18). Весьма трудоемкая работа по «добыванию» материала, кодированию и коррекции ошибок выполнена авторами этой публикации.

2.4 Электронные азбуки

Электронные азбуки знаменного распева строятся авторами на основе двознаменных певческих книг конца XVII – начала XVIII вв., представленных в формате «знамя–нота». Количество таких двознаменников невелико (порядка 10, см. [2]), однако именно они позволяют определить число, длительность и абсолютную высоту звуков, составляющих распев знамени. От певческих азбук Древней Руси (см. [3]) электронные азбуки отличаются тем, что указан источник, на основе которого составлена азбука, и по каждому знамени известна полная количественная информация, включающая частоту встречаемости знамени и всех его интерпретаций в каждом из 8 гласов. Различные интерпретации могут отличаться друг от друга интервально-ритмическими характеристиками и звуковысотной привязкой. Наличие количественной информации позволяет выделить доминирующую интерпретацию, если такая существует, а также интерпретации с аномально низкой частотой встречаемости (потенциально возможные ошибки). Электронные азбуки, созданные авторами на основе двознаменных Октоихов, описаны в [13, 14].

2.5 Обиходный звукоряд и система кодирования

Все знаменные песнопения записаны в диапазоне обиходного звукоряда (рис. 1). Он включает в себя ноты G, A, H (малой октавы), c, d, e, f, g, a, b (первой октавы), C, D

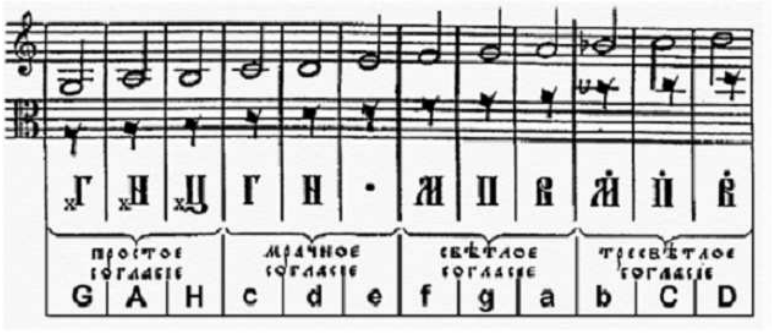


Рис. 1 Обиходный звукоряд

(второй октавы). В пометных рукописях ступеням звукоряда соответствуют обозначения , Ц, Г, Н, •, , П, , . Соответствующая знамени степенная помета указывает на наивысший звук в распеве знамени [2].

Особенности нотолинейной интерпретации знамен поясняются с помощью системы указательных помет: (или) — тихая, — ломка, — борзая, — ударка, — качка (или купно), — зевок, — равно. Знамена с указательной пометой и без нее трактуются как разные, поскольку могут иметь отличающиеся распевы. Длительности звуков обозначаются следующим образом: — 1 (целая), — 2 (половинная), — 4 (четвертная), — 8 (восьмая). Для обозначения высоты и длительности звука используем комбинацию буквы и цифры (например, $H4$ — это четвертная нота «си» малой октавы). При формальном сопоставлении нотолинейных цепочек такая комбинация буквы и цифры, характеризующая конкретный звук, рассматривается как один символ. Интервалы (число ступеней между высотами соседних звуков) кодируются целыми числами (1 — секунда, 2 — терция, 3 — кварта и т. д.) и сопровождаются знаком $(+)$ для восходящего движения и $(-)$ — для нисходящего. Например, $(3+)$ — это скачок на кварту вверх, $(2-)$ — на терцию вниз. При повторении звука на одной и той же высоте величина интервала условно обозначается $(0+)$. Заметим, что при таком кодировании не указывается число тонов и полутонов, содержащееся в интервалах, однако наличие нотолинейного текста дает возможность отличать, например, большую секунду $d4e4$ от малой секунды $e4f4$ или малую терцию $e2g2$ от большой $f2a2$. Знак $(*)$ используется в качестве разделителя между нотолинейными интерпретациями разных знамен, а $(\sim)$ — как символ эквивалентности, отделяющий знаменную цепочку от ее нотолинейного представления (например, запись $\sim d4c4d2 * c4e4$ означает, что цепочка из двух знамен, стоящая слева от $(\sim)$, интерпретируется в тексте двузнаменника, соответственно, цепочками из трех и двух нот, отделенными друг от друга знаком $(*)$).

2.6 Внутригласовые инварианты и квазиинварианты

Внутригласовые инварианты и квазиинварианты — специфический класс структурных единиц знаменного распева, описанных авторами в [9, 15]. Это повторяющиеся цепочки беспометных знамен относительно небольшой длины ($L = 1, 2, \ldots$), *однозначно* или *почти однозначно* (с допустимыми отклонениями) интерпретируемые в песнопениях конкретного гласа. Цепочки первого типа названы внутригласовыми инвариантами (ВИ), а второго — квазиинвариантами (КВИ). Места вхождения ВИ и КВИ в текст можно трактовать как своего рода «островки надежности», характеризующиеся существенным снижением уровня неоднозначности в интерпретации знамен. Предпосылкой к выявлению таких цепочек в тексте послужил анализ электронных азбук, построенных на основе двузнаменных Октоихов [13, 14]. Было замечено, что при всей вариативности знаменного распева в песнопениях каждого гласа существуют отдельные знамена, т. е. цепочки длины 1, не меняющие своей интерпретации. Обычно их доля не превышает 10% от размера алфавита. Менее очевидным было предположение о наличии однозначно интерпретируемых (в конкретном гласе) цепочек большей длины, составленных в общем случае из многозначных знамен. Однако это предположение подтвердилось. Словари ВИ и КВИ, построенные на основе двузнаменных Октоихов, легли в основу начальной версии алгоритма дешифровки беспометных песнопений [10].

Формальное определение КВИ требует уточнения понятия «допустимые отклонения». Здесь возможны различные варианты, требующие отдельной проверки. В начальной версии алгоритма [10] цепочки знамен трактуются как КВИ, если одна из ее интер-

претаций доминирует над остальными по частоте, а именно: выполняется соотношение $F_{\max}/F > 1/2$, где F — частота встречаемости цепочки в песнопениях гласа, представленных в обучающей подборке (двознаменнике), а $F_{\max}$ — максимальная из частот встречаемости разных интерпретаций этой цепочки. Доминирующая интерпретация, если она существует, трактуется в дальнейшем как значение КВИ. Наряду с указанным выше ограничением вводится ограничение на минимальную частоту встречаемости цепочки в песнопениях гласа: $F \geqslant F_{\min}$ (в [10] задавалось $F_{\min} = 3$).

Указанные ограничения приводят к тому, что не все знамена в дешифруемом песнопении оказываются покрытыми хотя бы одной цепочкой из словарей ВИ и КВИ, построенных на обучающем материале (имеет место «отказ»). Минимизировать количество отказов можно, либо видоизменив определение КВИ, либо объединяя подход, основанный на внутригласовых инвариантах, с какими-то другими, не использующими это понятие. В данной работе рассматриваются обе эти возможности. В частности, можно снять ограничение на доминирующую частоту $F_{\max} > F/2$ и всегда использовать в качестве КВИ (теперь уже *условного*) интерпретацию с максимальной частотой встречаемости, даже если не выполняется указанное выше соотношение. Оставляем лишь ограничение снизу, задав $F_{\min} = 2$ (цепочка должна быть повторяюшейся).

3 Сравнение алгоритмических подходов к дешифровке знаменных песнопений

На данный момент наиболее перспективными представляются два подхода. Первый основан на выявлении попевочной структуры песнопений, второй — на построении и использовании словарей инвариантов. Для обоих можно привести какие-то количественные показатели эффективности. Для подхода, намеченного в [8], пока не приведено никаких оценок эффективности, поэтому его не рассматриваем. Отдельные расшифровки, проводимые музыковедами, опираются на знания и интуицию исследователей и трудно формализуемы. Это ограничивает возможность использования накапливаемого в ходе таких работ опыта для целей *массовой* дешифровки.

3.1 Опора на попевочные структуры

Обращаясь к первому подходу (*попевочному*), отметим его естественность: именно различным комбинированием ограниченного числа гласовых попевок создаются новые песнопения. Имеется хороший задел для реализации подхода в виде многочисленных кокизников (сборников попевок). Схема дешифровки достаточно проста: нужно сформировать представительную подборку попевок по каждому гласу, записать их в формате *беспометное знамя – нота*, реализовать алгоритм отыскания знаменных попевочных цепочек в дешифруемом песнопении и заменять найденные попевки их нотолинейными эквивалентами. Однако на этом пути встречается много подводных камней, некоторые из них опишем на примере широко известной и достаточно представительной подборки попевок В. М. Металлова (см. подразд. 2.2).

1. Формального определения попевки не существует. На понятийном уровне это *устойчиво* повторяющийся мелодический оборот (см. подразд. 2.2). Понятие устойчивости, в свою очередь, допускает разные формализации, не приводящие к тождественным результатам. По этой причине любая систематизация попевок, включая приведенную в [12], будет неполной. Этим частично объясняются и большие разночтения в оценке объема попевочного фонда. В частности, применительно к подборке Металлова, считающейся достаточно представительной, неожиданные результаты дал экспери-

Таблица 1 Покрываемость двознаменника 619/647 попевками В. М. Металлова при поиске на точное совпадение

Гласы	Покрываемость (точное совпадение), %
1	31,1
2	30,3
3	22,4
4	24,8
5	30,8
6	23,7
7	32,1
8	29,4

мент с поиском попевок из этой подборки в нотолинейных текстах двознаменников. В табл. 1 приведены для примера показатели покрываемости разных гласов двознаменника 619/647 металловскими попевками (поиск осуществлялся на точное соответствие).

Здесь следует обратить внимание не столько на различия в показателях покрываемости в разных гласах, сколько на относительно невысокую степень покрытия. Применительно к дешифровке это означает, что лишь порядка 30% знаменного текста будет реконструировано с использованием подборки Металлова. Причин этому может быть несколько, одна из которых изложена ниже.

2. Попевки довольно вариативны. В слегка видоизмененной форме они все же могут быть обнаружены в нотолинейном тексте. Характер вариативности требует специального изучения. Авторы промоделировали простейший случай варьирования, предположив, что вариант может отличаться от базовой (металловской) попевки не более чем «m» допустимыми операциями, где в качестве элементарных операций выступают замена, вставка или устранение нотного знака в любой их комбинации. Параметр «m» выбирался равным 1 или 2 (для коротких попевок использовалось только значение

Таблица 2 Покрываемость двознаменника 619/647 попевками В. М. Металлова при наличии искажений

Гласы	Покрываемость, %	
	$m = 1$	$m = 2$
1	47,6	65
2	42,2	61,8
3	26,5	37.6
4	36	54,7
5	46	57,7
6	38,4	54,2
7	47	60,3
8	39	58

$m = 1$). Часть найденных вариантов отфильтровывалась экспертом вручную (недопустимой, например, считалась замена целой ноты в кадансовой структуре нотным знаком меньшей длительности или вставка, приводящая к немотивированному скачку звуковысотной линии, и т. п.). Покрываемость текста базовыми попевками и их вариантами увеличивалась (результаты для случаев $m = 1$ и 2 приведены в табл. 2). Приведенными значениями покрываемости и исчерпываются на данный момент возможности такого подхода. Увеличивать дальше значение m не имеет смысла ввиду лавинообразного нарастания ложных «попевок» и необходимости их отсеивания вручную. Более того, используемая модель хотя и фиксирует многочисленные (и подтверждаемые экспертом) случаи варьирования базовых попевок, тем не менее не в состоянии выявлять такие типы варьирования, как секвентный перенос попевки, вставку или устранение серии стопиц, реально встречающиеся на практике.

3. Значительные затруднения может вызвать необходимость представления попевок любой подборки в формате «беспометное знамя – нота», единственно пригодном для дешифровки беспометной нотации. Остается лишь гадать, почему подборка Металлова представлена только в нотолинейной форме, а попевки во многих кокизниках — только в знаменной. Предпринятые авторами попытки восстановления с помощью двузнаменников знаменных эквивалентов для нотолинейных попевок из подборки Металлова (см. [16]) показали, что: (а) далеко не для всех попевок подборки удается это сделать (нотолинейная цепочка из подборки не обнаруживается в нотолинейном тексте двузнаменника); (б) даже если она обнаруживается, ей не всегда соответствуют одинаковые цепочки знамен (вариативность имеет место и на уровне знамен). Приведем в качестве примера попевку «возмер» из седьмого гласа подборки Металлова. В кодировке авторов она записывается в виде: $e4d4e4f4g2f2e1$. Соответствующие этой попевке цепочки знамен в Октоихе 619/647 представлены ниже в виде выравнивания:

I	$e4d4$	$e4f4$	$g2$	$f2$	$e1$
II	$e4d4$	$e4f4$	$g2f2$		$e1$
III	$e4d4$	$e4f4g2$		$f2$	$e1$

Элементы варьирования проявляют себя в виде синонимичных замен в первом столбце выравнивания (знамена и имеют одинаковый распев), а также в последнем (здесь замену «статьи простой» на «статью мрачную» нельзя считать синонимичной, поскольку в звуковысотной иерархии вторая стоит выше первой, однако такие вещи встречаются, нередко приводя к ошибкам в дешифровке). Частым случаем варьирования является замена сложного знамени цепочкой из более простых, как это имеет место в строке «II» ($\sim$) и «III» ($\sim$). Возможно, вариативность на знаменном уровне и послужила причиной отсутствия знаменных эквивалентов у попевок из подборки Металлова.

Перечисленные выше замечания 1–3, связанные с реализацией первого подхода, затрудняют его использование на практике в качестве основного. Вызывает также сомнение подразумеваемый по умолчанию тезис о том, что песнопение может быть целиком покрыто только лишь комбинацией попевок. Определенные сложности представляет дешифров-

ка лиц и фит, хотя они встречаются не во всех песнопениях. Нельзя обойти молчанием тандемные повторы и регулярно встречающиеся на стыках между попевками короткие мелодические обороты без явных признаков каданса, подвода и других атрибутов попевочной структуры. Создается впечатление, что их используют «распевщики» для согласования текста с мелодией, а именно: для заполнения интервалов между предварительно расставленными по длине песнопения попевками, акцентирующими наиболее значимые фрагменты стихотворного текста. При дешифровке такие непервостепенные объекты, выполняющие связующие функции, дают много ошибок.

Завершая данный раздел, отметим, что описанный подход при всей своей привлекательности требует серьезной доработки. Однако не исключается возможность его использования уже на данном этапе в сочетании с каким-либо другим подходом, где бы они взаимно дополняли друг друга.

3.2 Опора на внутригласовые инварианты

Фактически речь идет о новом подходе к описанию всего корпуса знаменных песнопений с учетом системы осмогласия. Это описание — неполное, как и совокупность попевок, в том смысле, что по нему нельзя восстановить исходные данные (сами песнопения), но этого и не требуется. Словари ВИ и КВИ могут строиться либо по отдельным певческим книгам, либо по их совокупности. Выбор исходных данных важен. Забегая вперед, отметим, что лучше строить отдельные словари по каждому типу певческих книг (отдельно по Октоихам, отдельно по Ирмологиям и т. д.). Тогда при дешифровке конкретного песнопения, взятого, например, из Ирмология, лучших результатов достигнем, если воспользуемся словарями ВИ и КВИ, построенными именно по Ирмологиям.

Привлекательной стороной словарей ВИ и КВИ, рассматриваемых в качестве инструмента для дешифровки, является возможность *формального определения* составляющих эти словари цепочек знамен, их *интерпретируемость*, требуемый формат *представления* («беспометное знамя – нота»). Словари, построенные на основе трех двознаменных Октоихов, перечисленных в подразд. 2.3, достаточно представительны (порядка полутора тысяч цепочек знамен разной длины в каждом гласе песнопения). Это позволяет обеспечить достаточно высокую покрываемость дешифруемых песнопений.

Принципиальное отличие инвариантов от попевок — в диапазонах длин соответствующих им цепочек знамен. В соответствии с классификацией А. Н. Кручининой [12] уже кадансовые части попевок представлены цепочками из трех знамен. Им предшествуют «подводы», также состоящие из нескольких знамен. В то же время спектр длин цепочек из словарей ВИ и КВИ занимает диапазон от 1 до 10 и более знамен. Средние значения этого диапазона (5–8) примерно соответствуют длинам попевок. Более того, ВИ и КВИ длины 5 и выше часто сами являются попевками. Коротких ВИ и КВИ ($L = 1$–4) очень много, и именно за их счет обеспечивается более высокая (по сравнению с попевками) покрываемость дешифруемых песнопений. Короткие ВИ характеризуют наименее вариативные фрагменты попевок. Далеко не всегда при этом кадансовая часть попевки оказывается более устойчивой, чем подвод. Таким образом, ВИ и КВИ могут использоваться не только для дешифровки песнопений, но и для исследования структуры самих попевок.

Схема дешифровки сводится к поиску в песнопении фрагментов, совпадающих с цепочками знамен из словарей ВИ и КВИ, и приписыванию этим фрагментам певческих значений, зафиксированных в словарях. При этом возможны «*отказы*» (знамя не покрыто ни одной цепочкой из словарей ВИ и КВИ) и «*конфликты интересов*» (знамя входит в состав разных ВИ и КВИ, где ему приписываются неидентичные певческие значения).

Недостатком подхода является зависимость понятий ВИ и КВИ от объема и состава исходной (обучающей) подборки. При увеличении ее объема или изменении жанрового состава отдельные ВИ могут перейти в категорию КВИ, а некоторые КВИ перестают быть таковыми (нарушается условие доминирования: $F_{\max} > F/2$). Этот недостаток не оказывает существенного влияния на результаты дешифровки при достаточно представительной обучающей подборке. Другой недостаток связан с возможностью «отказов» (см. выше), что эквивалентно наличию пробелов в дешифровке. Для устранения этого недостатка требуется *интеграция* данного подхода с каким-то другим. В простейшем случае при наличии отказа можно приписать знамени соответствующую ему максимальную по частоте интерпретацию из электронной азбуки (в том же гласе). В значительном числе случаев эта интерпретация окажется правильной.

Более корректная и эффективная схема ориентируется на ближайших соседей знамени слева и справа. Пусть $T = t_1, t_2, \ldots, t_N$ — последовательность знамен в дешифруемом песнопении, и пусть знамя t_n $(2 \leqslant n \leqslant N-1)$ получило отказ в ходе дешифровки с использованием словарей ВИ и КВИ. Выделяем триграмму $t_{n-1}t_nt_{n+1}$ и находим максимальную по частоте интерпретацию этой триграммы в соответствующем гласе обучающей подборки. Певческое значение t_n заимствуем из этой триграммы. Для начального и конечного знамен используем триграммы $t_1t_2t_3$ и $t_{N-2}t_{N-1}t_N$ соответственно. Если указанные триграммы отсутствуют в обучающей подборке, восстанавливаем певческое значение t_n по электронной азбуке.

Начальная версия алгоритма дешифровки представлена в [10]. Текущая версия отличается от нее настройкой на конкретный тип певческой книги, из которой взяты дешифруемые песнопения (от этого зависит выбор словарей ВИ и КВИ). Реализована также описанная выше триграммная схема устранения отказов. Кроме того, в ситуациях, когда имеет место конфликт интересов и конкурирующие интерпретации набирают одинаковое число голосов, предпочтение отдается не любой из них (случайный шаг в начальной версии), а той, за которую «проголосовали» более длинные ВИ и КВИ. Результаты контрольных экспериментов с текущей версией описаны в разд. 5.

Оценка эффективности подхода, как и в [10], проводится на основе контрольного двознаменника. Показателем эффективности дешифровки одного песнопения служит отношение $k = n_+/N$ где N — длина песнопения (число знамен), а n_+ — количество знамен, получивших «правильную» (как в двознаменнике) интерпретацию. Показателем эффективности дешифровки группы песнопений, представляющих целый глас, служит отношение $\overline{k} = \overline{n}_+/\overline{N}$, где $\overline{N}$ — суммарная длина песнопений, а $\overline{n}_+$ — суммарное число правильно интерпретированных знамен в песнопениях гласа.

4 О возможности интеграции разных подходов

Примером такой интеграции уже является использование триграммной схемы устранения отказов в алгоритме дешифровки на основе словарей ВИ и КВИ. Эта схема, в принципе, может быть использована и самостоятельно для дешифровки всего песнопения. Показатели эффективности такой схемы колеблются в среднем от 50% до 60% для разных гласов.

Совместное использование внутригласовых инвариантов и попевок на данном этапе приводит к незначительному улучшению результатов по следующим причинам. Подборка Металлова в чистом виде, т. е. при $m = 0$, обеспечивает в среднем примерно 30%-ное покрытие песнопений разных гласов, что примерно в 2 раза уступает результатам дешифровки с использованием инвариантов. Надежда на то, что инварианты и попевки будут

«работать» как бы в противофазе, т. е. попевки будут попадать на участки концентрации отказов и исправлять ситуацию, также не оправдываются. Там, где значимо проявляют себя попевки, хорошо «работают» и инварианты, хотя бы в силу того, что инварианты значительной длины по большей части сами являются попевками. И, наконец, «конфликт интересов» может возникнуть не только между инвариантами, по-разному интерпретирующими одно и то же знамя, но и между металловской попевкой и нотолинейной компонентой контрольного двознаменника, которую мы считаем «правильной», равно как и попевку. Такая ситуация является реальной ввиду уже отмечавшейся выше вариативности попевок. Это может привести к тому, что попевка при наложении ее на уже дешифрованный с использованием ВИ и КВИ фрагмент песнопения может не только не улучшить результат, но даже ухудшить его (правильно реконструированное знамя получит другую интерпретацию).

Относительно небольшой положительный эффект, который достигается на данном этапе при совместном использовании инвариантов и попевок, может быть усилен при существенном (если не радикальном) «пополнении» подборки Металлова и обеспечении ее «знаменными эквивалентами» попевок (значительная часть подборки до сих пор их не имеет). Нуждается в уточнении и формализации и само понятие «попевки». Указанные аспекты требуют отдельной проработки.

5 Экспериментальные результаты

В данной работе для экспериментов использовали в основном двознаменные Октоихи и Ирмологий. В сумме по трем Октоихам на каждый глас приходилось порядка 70 песнопений, в Ирмологии — 35 песнопений. Результаты по двознаменнику «Праздники» не приводим, поскольку некоторые гласы (особенно седьмой) представлены в нем слишком малым числом песнопений. Исходные азбуки, а также словари ВИ и КВИ были построены: (1) на основе трех Октоихов (обобщенные); (2) на основе Ирмология. Результаты дешифровки в режиме скользящего контроля, описанного в [10], представлены в табл. 3 (отдельно для обоих типов певческих книг). Фиксировалось значение $F_{\min} = 2$, т. е. в формировании словарей ВИ и КВИ принимали участие только повторяющиеся цепочки знамен.

Прежде всего, обращает на себя внимание зависимость результатов от типа певческой книги: показатели для Октоихов заметно выше. Возможно, какое-то влияние оказывает различие в объемах материала, использовавшегося для формирования словарей ВИ и КВИ (суммарный объем Октоихов гораздо больше). Попытка использовать для дешиф-

Таблица 3 Доля правильно дешифрованных знамен $\overline{k}$ в режиме скользящего контроля для двух типов рукописей

Гласы	Октоихи	Ирмологий
1	0,767	0,720
2	0,666	0,628
3	0,731	0,618
4	0,723	0,675
5	0,730	0,691
6	0,662	0,569
7	0,763	0,812
8	0,745	0,632

ровки Ирмология словари ВИ и КВИ, построенные по более объемному материалу Октоихов, не увенчались успехом: результаты ухудшились примерно на 2%–5% в зависимости от гласа. То же самое наблюдалось и для тех гласов двознаменника «Праздники», которые были представлены достаточным количеством песнопений (~ 30). Отсюда и появилась рекомендация о необходимости согласования типов певческих книг дешифруемого песнопения и используемых словарей инвариантов.

Другая прослеживаемая по табл. 3 тенденция — снижение результатов дешифровки четных гласов (второго, шестого, а порой и восьмого) по сравнению с другими гласами. У нас нет пока сколь-нибудь убедительных толкований этой тенденции. Возможно, подсказкой послужит сопоставление гласов по попевочному составу. Гласы 2 и 6 охарактеризованы М. В. Бражниковым в [2, с. 201–202] как «богатые», а глас 7 — как «самый бедный».

В плане осознания дешифровочного потенциала разных подходов и возможностей их интеграции интерес представляет табл. 4, где условия дишифровки намеренно усложнены. Во-первых, используется пороговое значение $F_{\min} = 3$ (это ухудшает результат по сравнению со случаем $F_{\min} = 2$). Во-вторых, для дешифровки выбран Ирмологий, показавший худшие результаты в режиме скользящего контроля, чем Октоихи (см. табл. 3) и, наконец, использованы словари ВИ и КВИ, составленные по Октоихам, а не по Ирмологию, т. е. нарушается согласование по типу дешифруемых песнопений и используемых для этих целей словарей.

Реализованы 5 подходов, где три первых рассматриваются как независимые, а два последних являются их компиляцией.

Подход 1 определяет «нулевое» приближение, где используется только азбука. Каждому знамени в Ирмологии независимо от других присваивается певчекое значение, соответствующее наиболее часто встречающейся интерпретации этого знамени в азбуке, сформированной по Октоихам.

Подход 2 основан на использовании триграмм, т. е. учитываются связи между соседними знаменами. Триграммы выделяются из текста песнопения скользящим окном, сдвигающимся на каждом шаге на один символ вправо. Для дешифровки знамени t_n песнопения ($2 \leqslant n \leqslant N-1$, N — длина песнопения) выделяется триграмма $t_{n-1}t_nt_{n+1}$ и для нее находится максимальная по частоте встречаемости интерпретация в обучающей подборке (три Октоиха). Нотолинейную цепочку, соответствующую среднему знамени в этой триграмме, принимаем за значение t_n. Начальный и конечный символы дешифруем в соответствии с подходом 1.

Подход 3 — это дешифровка песнопений Ирмология с помощью словарей ВИ и КВИ, построенных на базе трех Октоихов (алгоритм описан в [10]).

Подход 4 — это сочетание подходов 3 и 1. За основу берется подход 3, а для устранения отказов используется подход 1, который может дать и ошибочный результат.

Подход 5 — это сочетание подходов 3 (ВИ и КВИ) и 2 (устранение отказов с использованием триграмм).

Таблицу 4 можно рассматривать как самое малое из того, что можно достигнуть с помощью описанных выше чисто статистических по своей сути подходов. Более оптимистичные и реальные оценки приведены в табл. 3. Удивляют, на первый взгляд, неожиданно высокие показатели, представленные в первом столбце. Это говорит о том, что среди всех возможных интерпретаций каждого знамени (за редким исключением) есть явно доминирующая по частоте, которая одна, сама по себе, может обеспечить значимый резуль-

Таблица 4 Доля правильно интерпретированных знамен $\overline{k}$ в различных схемах дешифровки Ирмология

Гласы	Подход 1	Подход 2	Подход 3	Подход 4	Подход 5
1	0,538	0,657	0,686	0,702	0,702
2	0,436	0,572	0,595	0,625	0,632
3	0,476	0,586	0,623	0,649	0,654
4	0,475	0,552	0,637	0,666	0,667
5	0,503	0,558	0,636	0,648	0,652
6	0,418	0,517	0,502	0,549	0,557
7	0,541	0,712	0,791	0,799	0,802
8	0,433	0,499	0,487	0,517	0,528

тат. Данное обстоятельство в определенной мере способствовало введению понятий ВИ и КВИ. Столбец 2 дает представление о силе связи между соседними символами. Столбец 3 указывает нижние пределы, которые при самых неблагоприятных условиях может обеспечить использование инвариантов. Незначительный прирост по сравнению со столбцом 2 обусловлен тем, что метод допускает отказы по некоторым (как правило, наиболее вариативным) позициям (в подходе 2 их нет). Следовательно, подход 3 нужно использовать в сочетании с каким-то другим, как минимум, для устранения отказов. Подходы 4 и 5 показывают, что использование простейших методов устранения отказов улучшает результат в среднем на 2%–5% в зависимости от гласа. Можно добавить также, что метод 3 в сочетании с попевками Металлова улучшает свой результат в среднем на 6%–7%, а метод 4 — всего на 1%–2% (результаты не приведены, поскольку носят предварительный характер).

Интеграция подхода 3 с 1 и 2 носила ограниченный характер, касалась только знамен, для которых имел место отказ. Предметом дальнейшего рассмотрения будет исследование возможности кооперации подходов 3 и 2 без указанного ограничения. Интерес представляет также исследование возможностей использования для дешифровки структур, названных авторами «условными инвариантами» в том смысле, что для них не является обязательным выполнение условия доминирования $F_{\max} > > F/2$.

Анализ ошибок дешифровки показывает, что подавляющая их часть связана с определением звуковысотной привязки знамен. Ошибки в восстановлении ритмической структуры встречаются редко. Повышенная концентрация ошибок наблюдается на стыках попевочных структур и в местах вхождения лиц и фит. Для последних, по-видимому, целесообразно строить отдельные словари, поскольку в разводах лиц и фит «рядовым» знаменем сохраняется еще элемент «тайнозамкненности».

Многие ошибки носят характер «допустимого варьирования» и тем самым фактически понижают приводимые в работе показатели качества. Например, знамя «стопица с очком» () обычно интерпретируется двумя четвертными в нисходящем движении (на ступень вниз). Именно на это и «настроены» словари инвариантов. Однако иногда, когда последний звук этого знамени совпадает по высоте с первым звуком следующего, делают скачок на две ступени вниз, что классифицируется как ошибка. Разбор подобных случаев «допустимого варьирования» и правильный их учет в показателях эффективности и дешифровки требует отдельного рассмотрения.

6 Заключение

Рассмотрены различные подходы к дешифровке древнерусских знаменных песнопений с привлечением компьютерно-ориентированных технологий. Область возможного использования ограничивается певческими книгами XVII в. и более позднего периода, допускающими и *беспометную* нотацию. Возможность применения рассматриваемых подходов для дешифровки беспометных песнопений XVI в. требует отдельного обоснования. Результат обработки — это последовательность знамен и предлагаемые их интерпретации, не исключающие, однако, наличие ошибок. Он может быть использован экспертом-медиевистом в качестве начального варианта, требующего последующей доработки.

Из рассмотренных вариантов решения предпочтение отдано методу, основанному на использовании словарей ВИ и КВИ, в сочетании с триграммной схемой обработки отказов в интерпретации, допускаемых основной процедурой. Эта схема интеграции разных подходов обеспечивает компромисс между качеством дешифровки и сложностью реализации. Понятия ВИ и КВИ введены авторами работы и оказались весьма продуктивными для дешифровки. Прослежена связь этих понятий с попевочной структурой песнопений. Словари ВИ и КВИ строятся на основе двознаменных певческих книг конца XVII – начала XVIII вв.

В серии экспериментов обоснована целесообразность построения и использования словарей для отдельных типов певческих книг (Октоихи, Ирмологии, Праздники и т. п.), а не их совокупности. При условии согласования типов дешифруемого песнопения и используемого словаря достигается 65%–80%-ная точность дешифровки по разным гласам в режиме скользящего контроля. Анализ ошибок показывает, что многие из них носят характер «допустимого варьирования».

Литература

[1] *Бражников М. В.* Пути развития и задачи расшифровки знаменного роспева XII–XVIII веков. — Л.—М.: Музыка, 1949. 104 с.

[2] *Бражников М. В.* Древнерусская теория музыки. По рукописным материалам XV–XVIII веков. — Л.: Музыка, 1972. 423 с.

[3] Певческие азбуки древней Руси / Публикации, перевод, предисловие и комментарии Д. Шабалина. — Кемерово: Кузбассвузиздат, 1991. 278 с.

[4] *Бражников М. В.* Новые памятники знаменного распева. — Л.: Музыка, 1967.

[5] *Смоляков Б. Г.* К проблеме расшифровки знаменной нотации // Вопросы теории музыки. — М., 1975. Вып. 3. С. 41–69.

[6] *Карастоянов Б. П.* К вопросу расшифровки крюковых певческих рукописей знаменного роспева // Musica Antique (Bydgoszcz), 1975. С. 485–504.

[7] *Школьник М. Г.* Проблемы реконструкции знаменного роспева XII–XVII веков: На материале византийского и древнерусского Ирмология. Автореф. дисс.. . . канд. иск. наук. — М., 1996. 23 с.

[8] *Даньшина М. В., Филиппович А. Ю.* Методика автоматизированной расшифровки знаменных песнопений // Вестник Московского государственного технического ун-та им. Н. Э. Баумана. Серия «Приборостроение», 2014. Вып. 4(97)С. 55–69.

[9] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Компьютерный поиск инвариантных структурных единиц знаменного распева // Проблемы музыкальной науки. Российский научный специализированный журнал, 2011. № 1(8). С. 20–24.

[10] *Бахмутова И. В., Гусев В. Д., Мирошниченко Л. А., Титкова Т. Н.* Параллельные тексты в задаче дешифровки древнерусских знаменных песнопений // Машинное обучение и анализ данных, 2015. Т. 1. №13. С. 1866–1876.

[11] *Металлов В. М.* Осмогласие знаменного распева (сборник нотолинейных попевок). — М., 1899. 50 с.

[12] *Кручинина А. Н.* Попевка в русской музыкальной теории XVII века. Автореф. дисс.. . . канд. иск. наук. — Л., 1979.

[13] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Электронная азбука знаменного распева: Предварительная версия // Вычислительные системы, 2005. Вып. 174. С. 29–53.

[14] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Создание электронной азбуки знаменного распева на основе анализа двознаменников // Древнерусское песнопение. Пути во времени, 2010. Вып. 4. С. 99–108.

[15] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Выявление инвариантов и квазиинвариантов знаменного распева с помощью билингв типа «знамя–нота» // Мат-лы Всерос. Конф. ЗОНТ-2013. — Новосибирск, 2013. Т. 1. С. 27–35.

[16] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Компьютерный анализ и восстановление знаменной составляющей подборки В. М. Металлова // Сибирский музыкальный альманах. — Новосибирск: Изд-во НГК, 2010. С. 66–85.

Поступила в редакцию 02.09.2016

Comparison and integration of approaches to deciphering Russian ancient chants

I. V. Bakhmutova, V. D. Gusev, L. A. Miroshnichenko, and T. N. Titkova

bakh@math.nsc.ru; gusev@math.nsc.ru; luba@math.nsc.ru; titkova@math.nsc.ru

S. L. Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences

4 Acad. Koptyug Str., Novosibirsk, Russia

The most promising approaches to the problem of noted reconstruction (deciphering) of ancient Russian hymnals of XVII–XVIII sc. written in neumatic (znamenny) form are considered. The approach, suggested by the authors, is preferable and it is based on use of invariants – neume chains that are interpreted with minimal level of ambiguity. The performance evaluations of different approaches are given. The possibility of their integration is studied.

Keywords: *dvoznamenniks; glasyi; popevki; the ancient Russian chants; deciphering; invariants; quasi-invariants*

DOI: 10.21469/22233792.2.4.03

References

[1] Brazhnikov, M. V. 1949. *Puti razvitiya i zadachi rasshifrovki znamennogo rospeva XII–XVIII vekov* [Development path and decoding tasks of XII–XVIII centuries Znamenny chant]. Leningrad–Moscow: Music. 104 p.

[2] Brazhnikov, M. V. 1972. *Drevnerusskaya teoriya muzyki* [Ancient Russian theory of music]. Leningrad: Music. 423 p.

[3] *Pevcheskie azbuki drevney Rusi* [Singing the alphabet ancient Russia]. Publikatsii, perevod, predislovie i commentarii D. Shabalina. [Publications, translation , introduction, and commentary D. Shabalina]. Kemerovo: Kuzbassvuzizdat. 278 p.

[4] Brazhnikov, M. V. 1967. *Novye pamyatniki znamennogo raspeva* [New monuments of znamenny chant]. Leningrad: Music.

[5] Smolyakov, B. G. 1975. K probleme rasshifrovki znamennoy notatsii [On the problem of Znamenny notation's decoding]. *Voprosy teorii muzyki* [Music Theory Questions] 3:41–69.

[6] Karastoyanov, B. P. 1975. K voprosu rasshifrovki kryukovykh pevcheskikh rukopisey znamennogo raspeva [On the question of deciphering the manuscripts of vocal hook znamenny chant]. *Musica Antique (Bydgoszcz).* 485–504.

[7] Shkolnik, M. G. 1996. Problemy rekonstruktsii znamennogo rospeva VII–XVII vekov: Na materiale vizantiyskogo i drevnerusskogo Irmologiya [Problems of reconstruction of Znamenny Chant of the XII–XVII centuries: On a material of Byzantine and Old Irmologion]. Ph.D. Thesis. Moscow. 23 p.

[8] Danshina, M. V., and A. Yu. Filippovich. 2014. Metodika avtomatizirovannoy rasshifrovki znamennykh pesnopeniy [The methodology of the automated decryption of znamenny chants]. *Vestnik Moskovskogo gosudarstvennogo tekhn. un-ta im. N. E. Baumana Ser. Priborostroenie* [Bull. N. E. Bauman Moscow State Technical University. Instrument ser.] 4(97): 55–69.

[9] Bakhmutova, I. V., V. D. Gusev, and T. N. Titkova. 2011. Komp'yuternyy poisk invariantnykh strukturnykh edinits znamennogo raspeva [Computerized search for the invariant structural elements of znamennyi chant]. *Music Scholarship* 1(8):20–24.

[10] Bakhmutova, I. V., V. D. Gusev, L. A. Miroshnichenko, and T. N. Titkova. 2015. Parallel'nye teksty v zadache deshifrovki drevnerusskikh znamennykh pesnopeniy [Parallel texts in the problem of deciphering of ancient Russian chant]. *Mach. Learn. Data Anal.* 1(13):1866–1876.

[11] Metallov, V. M. 1899. *Osmoglasie znamennogo raspeva (sbornik notolineynykh popevok)* [Osmoglasiya znamenny chant: Collection of notoline songs]. Moscow: Nauka. 50 p.

[12] Kruchinina, A. N. 1979. Popevka v russkoy musykal'noy teorii XVII veka [Popevka in Russian musical theory of the XVII century]. Ph.D. Thesis. Leningrad.

[13] Bakhmutova, I. V., V. D. Gusev, and T. N. Titkova. 2005. Elektronnaya azbuka znamennogo raspeva: Predvaritel'naya versiya [Electronic alphabeth of Russian chant: Previous version]. *Comput. Syst.* 174:29–53.

[14] Bakhmutova, I. V., V. D. Gusev, and T. N. Titkova. 2010. Sozdanie elektronnoy azbuki znamennogo raspeva na osnove analiza dvoznamennikov [Creation of an electronic alphabet znamenny chant based on the analysis of dvoznamennikov]. *Drevnerusskoe penie. Puti vo vremeni* [Ancient Chants. Time Path] 4:99–108.

[15] Bakhmutova, I. V., V. D. Gusev, and T. N. Titkova. 2013. Vyyavlenie invariantov i kvaziinvariantov znamennogo raspeva s pomoshch'yu bilingv tipa "znamya–nota" [Revelation of invariants and quasi-invariants of znamennyi chant using "neume-note" bilinguas]. *Russian Conference ZONT-2013 Proceedings.* Novosibirsk. 1:27–35.

[16] Bakhmutova, I. V., V. D. Gusev, and T. N. Titkova. 2010. Komp'yuternyy analiz i vosstanovlenie znamennoy sostavlyayushchey podborki V. M. Metallova [Computer analysis and restoration of znamenny component V. M. Metallov selections]. *Sib. Musical Almanach.* Novosibirsk: NGK. 66–85.

Received September 2, 2016

Вероятностная модель для сглаживания целевых метрик качества ранжирования*

Н. А. Волков[1,2]***, М. Е. Жуковский***[1,2]
nikitavolkov@yandex-team.ru; zhukmax@yandex-team.ru
[1]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9
[2]Яндекс, Россия, г. Москва, ул. Льва Толстого, 16

Задача информационного поиска — находить релевантные документы по запросу пользователя. Одной из важнейших задач информационного поиска является задача ранжирования результатов поиска. В настоящее время широко распространен подход получения функции релевантности с помощью методов машинного обучения на основе обучающей выборки. Для оценки качества принято использовать метрики качества, например, pFound. Однако большинство из них являются дискретными, что усложняет, например, отбор признаков. Цель данной работы — получить гладкий аналог дискретной метрики. Рассматриваются несколько вариантов определения гладкой метрики, лучший из которых экспериментально определяется по критериям гладкости и похожести на дискретную метрику. Критерий похожести, в отличие от критерия гладкости, численно выразить довольно трудно ввиду большого множества различных случаев поведения метрик, поэтому от численного описания пришлось отказаться. По результатам многочисленных экспериментов была найдена оптимальная модель гладкой метрики.

Ключевые слова: *задача ранжирования; метрики качества; сглаживание метрик; дискретные метрики; метрика pFound*

DOI: 10.21469/22233792.2.4.04

1 Введение

1.1 Задача ранжирования

Пусть D — некоторая коллекция текстовых документов, а Q — множество запросов. Обозначим D_q неупорядоченный набор документов, потенциально релевантных запросу $q \in Q$. Основная задача ранжирования — упорядочить документы внутри D_q по убыванию степени их релевантности запросу, т. е. более релевантные документы должны иметь более высокий ранг.

Если документ d_1 релевантнее документа d_2 по запросу $q \in Q$, то будем писать $(q, d_1) \triangleright (q, d_2)$. Таким образом определяется порядок на парах запрос–документ. Отметим, что этот порядок определен только внутри одного конкретного запроса.

Чтобы упорядочить пары запрос-документ, будем искать *функцию релевантности* $\alpha(q, d)$, которая удовлетворяет условию:

$$(q, d_1) \triangleright (q, d_2) \quad \Leftrightarrow \quad \alpha(q, d_1) > \alpha(q, d_2).$$

Разумеется, для того чтобы найти функцию релевантности, необходимо заранее организовать описанный порядок на парах запрос–документ. Для этой цели специалисты (работники коммерческих поисковых систем), называемые асессорами, для некоторых запросов размечают пары запрос–документ, определяя для них *оценку* релевантности $y(q, d)$.

*Работа выполнена при финансовой поддержке компании Яндекс.

На основе этой асессорской выборки и будем строить функцию α, которая приближает неизвестное $y(q,d)$ на всем множестве пар запрос–документ. Стоит также отметить, что в настоящее время в коммерческих поисковых системах процедура оценивания релевантности ассессорами достаточно хорошо формализована в виде списка многочисленных правил.

По этим данным составляется обучающая выборка $X_{\text{train}} = \{x_i = (q_i, d_i), y_i\}_{i=1}^{n}$, с помощью которой строится композиция деревьев $\alpha(x) = \sum_{i=1}^{N} a_i t_i(x)$, например λ-MART (см., например [1]).

1.2 Метрики качества

Различные ранжирования пар запрос–документ сравниваются при помощи метрик качества ранжирования. Для измерения качества ранжирования традиционно (традиция задана конференцией TREC — Text REtrieval Conference) используются меры точности, например метрики MAP (mean average precision), ERR (expected reciprocal rank) и NDCG (normalized discount cumulative gain).

Приведем описание метрики pFound, которая разработана в Яндексе на базе эмпирических исследований поведения пользователей [2, 3]. Она аппроксимирует вероятность того, что пользователь удовлетворится результатами поиска по запросу. В данном случае предполагается, что $y(q,d) \in \{0,1,2,3,4\}$. На основе экспериментов была получена зависимость от оценки асессоров вероятности $p(q,d)$ того, что пользователь удовлетворится документом d по запросу q. Оценкам 0, 1, 2, 3 и 4 соответствуют вероятности 0, 0,07, 0,14, 0,41 и 0,61. Также эмпирически было получено, что вероятность того, что пользователь прекратит поиск, равна $P_{\text{break}} = 0{,}15$ на каждом документе. Пусть набор документов $\{d_q^{(i)}\}_i$ упорядочен в соответствии со значением $\alpha(q,d)$. Тогда вероятность того, что пользователь при просмотре поисковой выдачи дойдет до i-го документа, равна

$$P_i = \left(1 - p\left(d_q^{(i-1)}\right)\right)\left(1 - P_{\text{break}}\right) P_{i-1}, \;\; P_1 = 1.$$

Окончательно:

$$\text{pFound}(n) = \sum_{i=1}^{n} P_i \; p\left(d_q^{(i)}\right).$$

2 Предпосылки к задаче

Выше был описан метод ранжирования, а также метрики качества. Для хорошего ранжирования используется набор признаков того, что данный документ релевантен запросу. При разработке нового признака возникает естественный вопрос, влияет ли он на качество функции релевантности.

2.1 Метод отбора признаков

Пусть U_0 — существующее множество признаков, а U_1 — оно же с добавлением новых признаков, т. е. $U_0 \subset U_1$. Разобьем выборку $X = \{x = (q,d), y\}$, размеченную асессорами, на две непересекающиеся части, которые обозначим X_{train} и X_{test}, и будем называть их обучающей и тестовой выборками соответственно. Обозначим F_{X_0,U_i} модель, обученную по выборке X_{train} на множестве признаков U_i, т. е. по векторам $x = (f(q,d))_{f \in U_i}$. Значением метрики MSE (mean-squared error) по множеству признаков U_i будем называть величину

$$\text{MSE}\left(U_i\right) = \sqrt{\frac{1}{|X_{\text{test}}|} \sum_{x \in X_{\text{test}}} \left(F_{X_0,U_i}(x) - p(x)\right)^2},$$

где $p(q, d)$ — вероятность того, что пользователь удовлетворится документом d по запросу q.

Для вынесения решения о том, что добавление признаков дает прирост в качестве, можно, например, проверить статистическую гипотезу:

$$H_0 : \mathrm{MSE}\,(U_1) - \mathrm{MSE}\,(U_0) > 0.$$

Проверка такой гипотезы происходит с помощью N-кратного разбиения вида $X = X^n_{\text{train}} \cup X^n_{\text{test}}$, $n = 1, \ldots, N$, и подсчета для каждого разбиения числа $\mathrm{MSE}^n(U_i)$ для наборов признаков U_0 и U_1, получая тем самым парную выборку. По полученной выборке строится тест Уилкоксона [4].

Недостатком данного подхода является тот факт, что метрика MSE в силу своего определения предназначена для оценки точности предсказания абсолютных значений функций релевантности $\alpha(q, d)$, а не порядка, который они определяют.

2.2 Дискретность метрик

Выше были описаны недостатки использования метрики MSE. Но почему же тогда она используется при отборе признаков, а не специальная метрика, предназначенная для оценки качества ранжирования, например pFound? Дело в том, что такие метрики дискретны, в силу того что они зависят только от порядка. Поясним эту проблему рассмотрением двух случаев.

Случай 1. Пусть α' и α'' — некоторые функции релевантности, причем их значения на документах d_1 и d_2 по запросу q отличаются не сильно, но так, что порядок этих документов меняется. Формально это можно записать следующими условиями:

$$\alpha'(q, d_1) = \alpha'(q, d_2) + \varepsilon_1\,;$$

$$\alpha''(q, d_1) = \alpha''(q, d_2) - \varepsilon_2\,,$$

где $\varepsilon_1 > 0$ и $\varepsilon_2 > 0$ малы. Поскольку порядок меняется, то меняется и значение метрики.

Случай 2. Пусть теперь α' и α'' — функции релевантности, построенные так, что порядок документов d_1 и d_2 по запросу q не меняется. Однако разница значений функций релевантности для этих документов меняется сильно. Формально это можно записать следующими условиями:

$$\alpha'(q, d_1) = \alpha'(q, d_2) + \beta_1\,;$$
$$\alpha''(q, d_1) = \alpha''(q, d_2) + \beta_2\,,$$

где $\beta_2 > 0$ и $\beta_1 > 0$, но при этом $\beta_1 - \beta_2$ велико. Поскольку порядок не меняется, то не меняется и значение метрики.

Получаем, что в одном случае значения функций релевантности меняются не сильно, но это влечет изменение порядка, а значит, и изменение значения метрики, а в другом случае значения функций релевантности меняются сильно, но при этом такое изменение не влечет изменение порядка, а значит, и не меняется значение метрики. Теперь ясно, почему хорошие метрики, такие как NDCG и pFound, являются дискретными.

3 Постановка задачи и первые попытки решения

В предыдущем разделе были описаны минусы дискретных метрик, которыми являются все основные метрики качества ранжирования. В связи с этим возникает естественное желание получить хорошую гладкую метрику. На самом деле, получить гладкую метрику

может быть не так сложно. Гораздо сложнее добиться того, чтобы она была действительно хорошей в следующем смысле: она должна быть похожа на сглаживаемую дискретную метрику.

В силу всего вышесказанного формулируем задачу следующим образом: *получить гладкую метрику, похожую на хорошую дискретную метрику.* Стоит отметить, что данная формулировка не дает четкого определения слов «гладкая» и «похожая». Они будут пояснены при проведении экспериментов позже.

Существующие способы сглаживания метрик и их недостатки

В статьях [5, 6] рассматривается задача замены дискретных метрик некоторыми гладкими функциями потерь, которые называются суррогатными. К сожалению, такой подход не дает решения данной задачи: ведь он основан только на совпадении решений задач оптимизации дискретной метрики и суррогатной функции потерь. Мы же требуем «похожесть» гладкой метрики в гораздо более широком смысле. В статье [7] решается задача непосредственного сглаживания метрики ERR.

Все известные авторам попытки разработки желаемой метрики основывались на введении некоторой вероятностной модели так, чтобы ранжирование получалось случайным. Тогда в соответствии с этой вероятностной моделью эксперименты можно проводить несколько раз и усреднять результаты, получая тем самым сглаженное значение метрики.

Первые попытки введения вероятностной модели использовали метод перестановки документов в запросе на основе модели Plackett–Luce [8]. Случайные перестановки создавались с помощью некоторого распределения с учетом весов документов — значений функции релевантности на текущей итерации. В этом случае числа, которые выдает модель, определяют не релевантность документа, а вероятность того, что этот документ может быть более релевантным данному запросу. Само ранжирование генерируется с помощью полученного распределения. Таким образом, ожидаемый pFound можно представить как среднее значение по всем полученным перестановкам документов, отвечающих запросу.

Однако от этого способа пришлось отказаться, так как в таком случае для подсчета нужно выполнить $N!$ перестановок, если рассматривать только N наиболее релевантных документов, что является достаточно долгой процедурой. Если проводить оценку методом Монте Карло, то полученная оценка оказывалась слишком неточной — при одинаковых экспериментах значения метрики различались слишком сильно, из-за чего случайные факторы проходили тест отбора признаков.

4 Метрика EpFound

В предыдущем разделе были перечислены недостатки введения вероятностной модели на результатах ранжирования. Из-за этих недостатков был выбран другой метод введения вероятностной структуры — вероятностная модель вводится на самих документах. Это в некотором смысле естественно, так как, например, при смене базы D значения признаков могут несколько измениться.

4.1 Базовая модель

Пусть X — некоторое множество пар запрос–документ. Будем считать его случайным с неизвестным распределением P, из которого сгенерировано именно это множество. Это распределение будем называть истинным распределением множества X. Например, P — это распределение множества пар запрос–документ для поисковой машины yandex.ru, а для yandex.com распределение может быть другим.

Разделим множество X на две непересекающиеся части X_{train} и X_{test}, которые будем называть обучающей и тестовой выборками соответственно. Обозначим $\text{pFound}(F(X_{\text{train}}), X_{\text{test}})$ значение pFound на выборке X_{test} по формуле F, обученной по выборке X_{train}. Для того чтобы избавиться от дискретности метрики, можно посчитать ее математическое ожидание $\mathsf{E}_{\mathsf{P}}\text{pFound}(F(X_{\text{train}}), X_{\text{test}})$, где E_{P} — математическое ожидание в предположении, что множества имеют распределение P.

Посчитать это математическое ожидание нам не удастся по двум причинам. Во-первых, распределение может быть чересчур сложным. Однако эту проблему можно решить, сделав оценку математического ожидания с помощью метода Монте Карло

$$\widehat{\mathsf{E}}_{\mathsf{P}}\text{pFound}\left(F(X_{\text{train}}), X_{\text{test}}\right) = \frac{1}{M}\sum_{m=1}^{M}\text{pFound}\left(F(\xi^m_{\text{train}}), \xi^m_{\text{test}}\right),$$

где $(\xi^m_{\text{train}}, \xi^m_{\text{test}})$ сгенерировано из распределения P. Вторая проблема заключается в том, что само истинное распределение P неизвестно. Далее в работе будем пытаться построить оценку P^* для распределения P.

Допустим, мы уже построили некоторую оценку P^* для распределения P. Тогда мы можем оценить неизвестное нам математическое ожидание $\mathsf{E}_{\mathsf{P}}\text{pFound}(F(X_{\text{train}}), X_{\text{test}})$ по распределению P математическим ожиданием $\mathsf{E}_{\mathsf{P}^*}\text{pFound}(F(X_{\text{train}}), X_{\text{test}})$ по распределению P^*, которое, в свою очередь, оценим с помощью метода Монте Карло:

$$\widehat{\mathsf{E}}_{\mathsf{P}^*}\text{pFound}(F(X_{\text{train}}), X_{\text{test}}) = \frac{1}{M}\sum_{m=1}^{M}\text{pFound}(F(\xi^m_{\text{train}}), \xi^m_{\text{test}}),$$

где $(\xi^m_{\text{train}}, \xi^m_{\text{test}})$ сгенерировано из распределения P^*.

Теперь займемся построением оценки P^* распределения P. Для этого примем несколько предположений о распределении P^*, которые описаны ниже.

Обозначим $x = (q, d)$ некоторую пару запрос–документ. Будем обозначать также признаки пары запрос–документ x как f_k. Для каждого признака f_k определим соответствующие ей бинарные признаки g_{kj} по правилу $g_{kj} = I\{f_k(x) > b_{kj}\}$, где $\{b_{k1}, \ldots, b_{kB_k}\}$ — упорядоченный по возрастанию набор точек разбиения, $B_k \geqslant 0$.

Предположение 1. Будем искать оценку P^*, предполагая независимость в совокупности исходных признаков документа.

Предположение 2. Для облегчения процедуры получения оценки P^* примем также заведомо неверный факт о независимости в совокупности бинарных компонент $g_{k1}, \ldots, g_{kn}$ одного небинарного признака f_k. Заметим, что при таком предположении нарушается свойство бинарных признаков, которое будем называть *свойством монотонности*. Это свойство заключается в том, что для любых $a \neq (0, \ldots, 0, 1, \ldots, 1)$ выполнено равенство $\mathsf{P}\left(g_{k1}(x) = a_1, \ldots, g_{kn}(x) = a_n\right) = 0$. Однако в предположении независимости бинарных компонент одного небинарного признака получаем:

$$\mathsf{P}^*\left(g_{k1}(x) = a_1, \ldots, g_{kn}(x) = a_n\right) = \mathsf{P}^*\left(g_{k1}(x) = a_1\right)\cdots\mathsf{P}^*\left(g_{kn}(x) = a_n\right),$$

где правая часть не равна нулю, например, если каждый бинарный признак не является вырожденной случайной величиной.

Создание распределения

Определим окрестность $U_N(x)$ — множество из N ближайших соседей пар запрос–документ x в бинаризованном пространстве по L_1-метрике, которая определяется как

$$\rho(x, x') = \sum_k \sum_j |g_{kj}(x) - g_{kj}(x')| .$$

По умолчанию сама пара запрос–документ x не включается в окрестность $U_N(x)$.

Обозначим $p_{kj}(x) = \mathsf{P}(g_{kj}(x) = 1)$. В силу указанных выше предположений набор оценок $\{p^*_{kj}(x)\}_{x,k,j}$ таких чисел задает оценку распределения P^* для бинаризованного множества. Определим их с помощью пар запрос–документ из окрестности $U_N(x)$:

$$p^*_{kj}(x) = \mathsf{P}^*(g_{kj}(x) = 1) = \frac{1}{N} \sum_{x' \in U_N(x)} g_{kj}(x').$$

Эта оценка является оценкой максимального правдоподобия для $p_{kj}(x)$ по выборке документов из окрестности $U_N(x)$.

Генерация множеств

Пусть $X = \{g_{kj}(x)\}_{x,k,j}$ — некоторое бинарное множество и $\mathsf{P}^*(X) = \{p^*_{kj}(x)\}_{x,k,j}$ — оценка распределения множества. Задав распределение, можно сгенерировать из него новые множества. Новое случайное множества $\xi = \{\xi_{kj}(x)\}_{x,k,j}$ генерируется по правилу $\xi_{kj}(x) \sim \mathrm{Bern}\left(p^*_{kj}(x)\right)$. Обозначим эту процедуру $\xi \sim \mathrm{Bern}\left(\mathsf{P}^*(X)\right)$.

Определение EpFound

Как и раньше, будем обозначать X_{train} и X_{test} обучающую и тестовую выборки, и $\mathrm{pFound}\left(F(X_{\text{train}}), X_{\text{test}}\right)$ — значение pFound на X_{test} по формуле F, обученной на X_{train}. Как было сказано выше, будем определять EpFound по правилу:

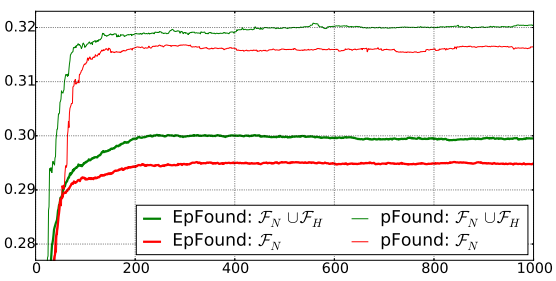
$$\mathrm{EpFound} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{pFound}\left(F(\xi^m_{\text{train}}), \xi^m_{\text{test}}\right),$$

где $\xi^m_{\text{train}} \sim \mathrm{Bern}\left(\mathsf{P}^*(X_{\text{train}})\right), \xi^m_{\text{test}} \sim \mathrm{Bern}\left(P^*(X_{\text{test}})\right)$.

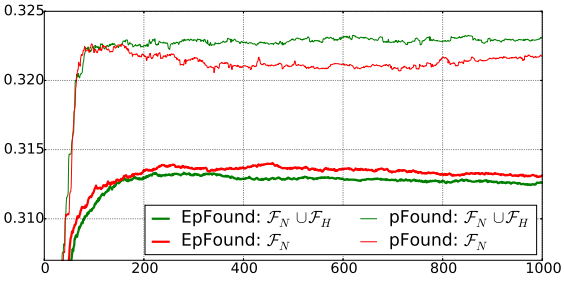
4.2 Проблемы базовой модели

Определенный выше вариант определения EpFound имеет несколько проблем. Самой основной из них является проблема различного поведения pFound и EpFound. Эта проблема может проявляться различными способами. Самый частый случай заключается в различном поведении этих метрик при использовании хостовых признаков и при их отсутствии. Хостовыми называем те признаки, которые полностью определяются хостом страницы и одинаковы для всех страниц данного хоста.

Пусть $\mathcal{F}_N$ и $\mathcal{F}_H$ — множества нехостовых и хостовых признаков соответственно. Одно из проявлений упомянутой проблемы заключается в том, что если pFound, посчитанный при использовании множеств с признаками $\mathcal{F}_N$, принимает значения *меньшие*, чем при использовании таких же множеств, но с признаками $\mathcal{F}_N \cup \mathcal{F}_H$, то EpFound наоборот, при использовании множеств с признаками $\mathcal{F}_N$ может принимать *большие* значения, чем при использовании таких же множеств, но с признаками $\mathcal{F}_N \cup \mathcal{F}_H$. Эта проблема показана на рис. 1, *а*, на котором видно, что относительный порядок для этих двух наборов признаков меняется при замене pFound на EpFound.



(*а*) Набор данных 1



(*б*) Набор данных 2

Рис. 1 График значения метрик pFound и EpFound в зависимости от количества деревьев в композиции; усреднение по 7 итерациям

Другое проявление проблемы хостовых признаков заключается в следующем. В некоторые моменты времени при построении композиции деревьев значение pFound возрастает с ростом количества деревьев в композиции, если используются признаки $\mathcal{F}_N \cup \mathcal{F}_H$, и убывает, если используются только признаки $\mathcal{F}_N$. EpFound же в таких случаях вообще практически не меняет значения, хотя при этом и принимает большие значения при использовании хостовых признаков, чем без них.

Эта проблема показана на рис. 1, *б*. Видно, что значение pFound после добавления 350-го дерева до добавления 400-го дерева в композицию деревьев, которые составляют обученную формулу, немного возрастает при использовании хостовых признаков и убывает без их использования. Однако у EpFound такого эффекта нет.

4.3 Способы решения проблем

Варианты усреднения

Выше был рассмотрен только один случай усреднения, а именно: мы определяли EpFound по правилу:

$$\text{EpFound} = \frac{1}{M}\sum_{m=1}^{M} \text{pFound}\left(F\left(\xi^m_{\text{train}}\right), \xi^m_{\text{test}}\right),$$

где $\xi^m_{\text{train}} \sim \text{Bern}\left(P^*(X_{\text{train}})\right)$; $\xi^m_{\text{test}} \sim \text{Bern}\left(P^*(X_{\text{test}})\right)$; X_{train} и X_{test} — обучающая и тестовая выборки; $\text{pFound}\left(F(X_{\text{train}}), X_{\text{test}}\right)$ — значение pFound на X_{test} по формуле F, обученной на X_{train}. Тем самым при усреднении генерируются как обучающая, так и тестовая выборки. Этот вариант будем называть первой моделью и обозначать его EpFound_1.

Здесь возникает вопрос: нужно ли генерировать как обучающую, так и тестовую выборки? Может быть, стоит генерировать только одну из них, а в качестве второй брать исходную? Во втором варианте будем генерировать только тестовые выборки, тем самым обучая формулу только один раз:

$$\text{EpFound}_2 = \frac{1}{M}\sum_{m=1}^{M} \text{pFound}\left(F(X_{\text{train}}), \xi^m_{\text{test}}\right),$$

где $\xi^m_{\text{test}} \sim \text{Bern}\left(P^*(X_{\text{test}})\right)$.

В третьем варианте будем генерировать только обучающие выборки, тем самым применяя одну тестовую выборку к набору обученных формул:

$$\mathrm{EpFound}_3 = \frac{1}{M}\sum_{m=1}^{M} \mathrm{pFound}\left(F(\xi^m_{\mathrm{train}}), X_{\mathrm{test}}\right) ,$$

где $\xi^m_{\mathrm{train}} \sim \mathrm{Bern}\left(\mathsf{P}^*(X_{\mathrm{train}})\right)$.

Формально эти формулы можно получить, взяв в качестве оценки распределения P^* вырожденное распределение для обучающей или тестовой выборки.

Окрестности

В базовом варианте мы предполагали, что окрестность $U_N(x)$ — множество из N ближайших соседей пар запрос–документ x в бинаризованном пространстве по L_1-метрике, причем сама пара запрос–документ x не включается в окрестность. Возможен также вариант, при котором пара запрос–документ x включается в окрестность $U_N(x)$ наравне с остальными документами.

Отдельно будем рассматривать вариант, при котором сама пара запрос–документ x не включается в окрестность $U_N(x)$, но в вычислении вероятностей берется с весом $w \in (0, 1)$:

$$p^*_{kj}(x) = w\ g_{kj}(x) + \frac{1-w}{N-1}\sum_{x' \in U_{N-1}(x)} g_{kj}(x').$$

Биномиальная модель

Также рассматривалась модель построения оценки P^*, для которой выполняется свойство монотонности. В этом случае мы генерировали не сами бинарные признаки, а номер последнего бинарного признака для данного исходного признака, которая принимает значение ноль. Распределение этой величины считалось биномиальным, параметр которого оценивался по окрестности аналогичным способом. Однако в данных экспериментах такой метод не привел к успеху. Условия экспериментов аналогичны описанным ниже.

5 Эксперименты

В предыдущих разделах была сформулирована исследуемая задача поиска гладкой метрики, рассказано, почему нас не удовлетворяют существующие способы сглаживания метрик, а также приведены теоретические описания вариантов новой сглаженной метрики. В данном разделе опишем эксперименты, которые были проведены для поиска наилучшего варианта из предложенных.

5.1 Описание данных

Данные для проведения экспериментов были получены с помощью поисковой машины yandex.ru, а также асессоров, которые разметили документы. Данные содержат 100 026 пар запрос–документ, из которых 75 021 пар составляли обучающую выборку, а остальные 25 005 пар — тестовую выборку. Обучающая выборка при этом состояла из 7181 запросов, а тестовая — из 2035 запросов. Разбиение выборок проводилось по запросам случайным образом, т. е. для каждого запроса все пары запрос–документ с этим запросом попадали либо в обучающую выборку, либо в тестовую.

В данных каждая пара запрос–документ описывается меткой релевантности, уникальным идентификатором запроса, а также набором признаков. Все признаки либо бинарные —- принимают значения из $\{0, 1\}$, либо непрерывные, значения которых нормированы на интервал $[0, 1]$. Метки релевантности принимают значения во множестве $\{0, 1, 2, 3, 4\}$.

Эксперименты проводили на трех различных множествах признаков. Первые два были получены случайным выбором 100 нехостовых признаков и 30 хостовых, третий — случайным выбором 100 нехостовых признаков и 14 хостовых.

5.2 Как измерялось качество сглаживания

Напомним сначала основную цель исследования: *получить гладкую метрику, похожую на* pFound. Как было отмечено выше, данная формулировка не дает четкого определения слов «гладкая» и «похожая». Теперь пришло время несколько формализовать данные понятия.

Критерий гладкости. Усредненная метрика имеет гладкий график зависимости ее значения от количества деревьев. Степень гладкости удалось выразить численно. Опишем процедуру подсчета степени гладкости зависимости.

Пусть $a = (a_1, \ldots, a_n)$ — некоторая зависимость. Определим *сглаженное значение* $\widehat{a}_i$ как значение линейной регрессии в точке i, построенной по точкам $(i - r, a_{i-r}), \ldots, (i + r, a_{i+r})$, причем из набора исключены s минимальных и s максимальных значений. Тогда *степенью гладкости* называется величина

$$S(a) = \frac{10^{-7}}{\mathrm{MSE}\,(a, \widehat{a})}.$$

В данных экспериментах использовались $r = 20$ и $s = 5$. Домножение на константу 10^{-7} сделано для того, чтобы значения гладкости получались приятными на глаз, в данных экспериментах это обычно значения от 0 до 100. Гладкость тем выше, чем больше значение степени гладкости для зависимости.

Критерий похожести. Усредненная метрика ведет себя «похоже» на *pFound*. К сожалению, данный критерий формализовать достаточно сложно. Поясним немного эту проблему.

Любая формализация данного критерия должна предполагать сравнение изменения трендов значений pFound и EpFound в зависимости от количества деревьев. Например, если pFound начинает расти, то и EpFound должен расти, и наоборот. Однако в случае, представленном на рис. 1, мы вовсе не требуем такую зависимость. Дело в том, что данное поведение можно еще интерпретировать как увеличение разности значения pFound при использовании хостовых признаков и без них. Такой же эффект мы хотим получить от EpFound. Это означает, что нас даже устроит случай, при котором абсолютные значения EpFound уменьшаются, но их разность возрастает.

Подобных частных случаев похожего поведения можно придумать достаточно много, поэтому какие-либо попытки формализации данного критерия скорее всего будут заточены под такие частные случаи.

5.3 Выбор лучшей модели

По критерию похожести лучше всего оказывается *простая модель с третьим типом усреднения*, в которой генерируется только обучающая выборка. Однако она лишь немного уступает упомянутой выше биномиальной модели по критерию гладкости и имеет значения гладкости от 20 до 25, в то время как биномиальная модель имеет значения гладкости от 35 до 40. Поскольку большей гладкости можно добиться увеличением количества формул при усреднении, данный недостаток можно считать незначительным. В свою очередь биномиальная модель оказывается намного хуже по критерию похожести. Таким образом, можно заключить, что эта простая модель с третьим типом усреднения является наилучшей. Теперь же нужно для нее подобрать оптимальное значение веса.

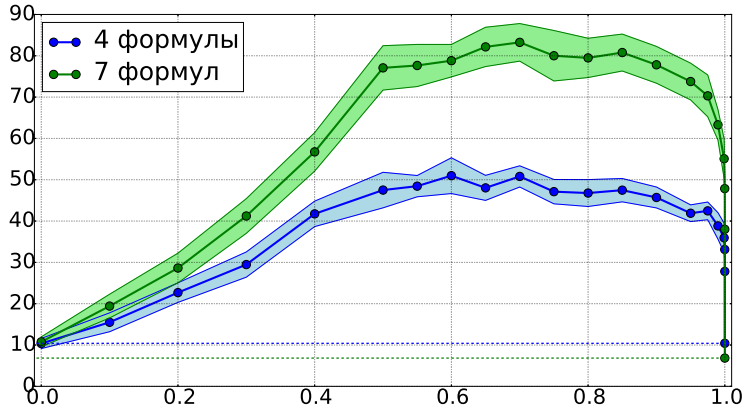


Рис. 2 График доверительных интервалов уровня доверия 0,95 для среднего значения степени гладкости

Было замечено (результаты экспериментов не приводятся, поскольку они заняли бы много места), что критерии гладкости и похожести часто дают схожие результаты, т. е. если одна модель хорошая с точки зрения критерия похожести, то она будет хорошей и по критерию гладкости, и наоборот. Таким образом, для нахождения его веса можно найти оптимальное значение по критерию гладкости, а затем проверить по критерию похожести.

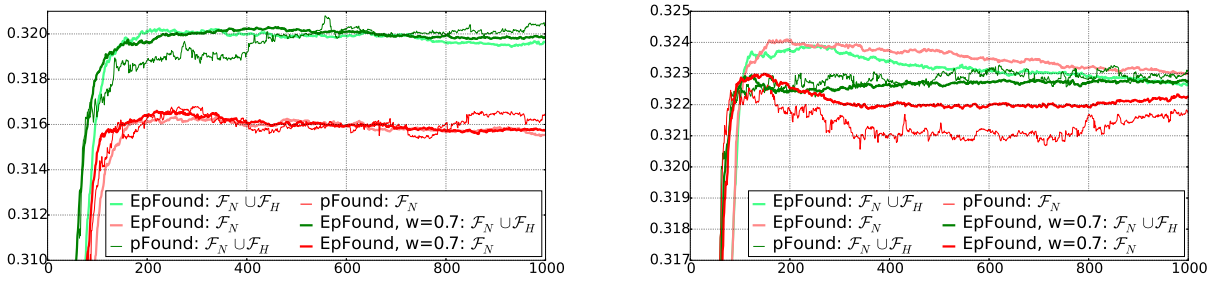
Для нахождения оптимального значения веса по критерию гладкости была проделана следующая процедура для некоторых значений весов. Для значения веса w было посчитано 10 значений степени гладкости EpFound на множестве 2 без использования хостовых признаков и 10 значений с их использованием. Полученные значения являются выборкой из некоторого неизвестного распределения. Из предположения, что эти значения имеют нормальное распределение, по ним был построен доверительный интервал уровня доверия 0,95 для параметра сдвига. График полученных доверительных интервалов приведен на рис. 2.

Доверительные интервалы были построены только для множества 2. Дело в том, что такая процедура является достаточно ресурсоемкой задачей. Для построения этого графика потребовалось обучить 2942 моделей, каждая из которых представляет из себя композицию из 1000 деревьев.

Из графика видно, что степень гладкости монотонно возрастает при возрастании веса от 0 до 0,5. Для значении веса от 0,5 до 0,9 значение степени гладкости меняется несильно. При приближении веса к 1 происходит резкое уменьшение значений степени гладкости. Вес $w = 1$ соответствует обычному pFound. Из графика видно, что оптимальное значение веса находится в окрестности 0,7, поэтому будем считать оптимальным вес $w = 0{,}7$.

Посмотрим теперь, насколько хорошо модель с весом $w = 0{,}7$ удовлетворяет критерию похожести.

На множестве 1 значения и поведения метрик EpFound и pFound практически совпадают, в частности схожа динамика перед добавлением 400-го дерева. Значение степени гладкости равно 77,4. Некоторое отличие значений pFound от EpFound наблюдается только в случае использования хостовых признаков от добавления 100-го дерева до добавления 400-го дерева. Однако поскольку до 100-го дерева и после 400-го дерева значения



(*а*) Набор данных 1 (*б*) Набор данных 2

Рис. 3 График значения метрик pFound и EpFound в зависимости от количества деревьев в композиции; усреднение по 7 итерациям, модель усреднения 3, вес 0,7

обоих метрик практически совпадают, можно сделать предположение о некотором недостатке самого pFound, который устраняется с помощью метрики EpFound.

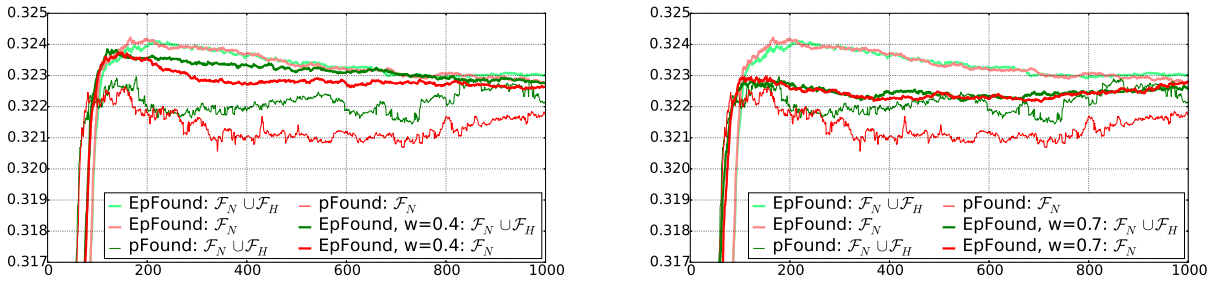
На множестве 2 поведение метрики EpFound достаточно хорошо повторяет поведение метрики pFound. При использовании модели с оптимальным весом значения метрики EpFound возрастают при достаточно большом количестве деревьев при использовании только нехостовых признаков, что полностью соответствует поведению метрики pFound (рис. 3). Значение степени гладкости равно 75,9.

Несмотря на то что метрика EpFound с оптимальным весом оказалась лучше других моделей по всем критериям на множествах 1 и 2, на множестве 3 она уступает моделям с другими весами. Дело в том, что ее значения при использовании хостовых признаков и при их отсутствии практически совпадают. Лучшим же по критерию похожести оказывается модель с весом $w = 0{,}4$. Значения степеней гладкости для весов 0,4 и 0,7 равны соответственно 60,0 и 77,4. Такая модель на множествах 1 и 2 оказывается хуже модели с весом $w = 0{,}7$ по критерию похожести.

6 Выводы и планы на будущее

На основании всех проведенных экспериментов можно сделать следующие выводы.

1. Была поставлена задача получить гладкую метрику, похожую на pFound. Для ее решения были предложены несколько методов, которые различаются способом генерации выборки, типом усреднения, а также видом рассматриваемых окрестностей точек. По совокупности критериев чаще всего лучшей оказывается простая модель EpFound с третьим типом усреднения, в которой генерируется только обучающая выборка. В связи с этим ее можно считать метрикой, которая лучше всего решает поставленную задачу.
2. Вес точки при подсчете вероятности является гиперпараметром модели. Для разных данных по различным критериям может быть свое значение оптимального веса. Наилучшие результаты получаются при значении веса от 0,4 до 0,9.
3. Качество оптимальной метрики как по критерию похожести, так и по критерию гладкости, сильно зависит от используемого множества.
4. Биномиальная модель, в которой сгенерированные значения бинарных признаков удовлетворяют свойству монотонности, имеет преимущество по критерию гладкости, но сильно проигрывает по критерию похожести. В связи с этим биномиальная модель не является оптимальной.



(*а*) Вес 0.4 (*б*) Вес 0.7

Рис. 4 График значения метрик pFound и EpFound в зависимости от количества деревьев в композиции, усреднение по 7 итерациям, модель усреднения 3, набор данных 3

5. Вторая модель усреднения, в которой генерируется только тестовая выборка, чаще оказывается хуже остальных, в особенности по критерию похожести. Поэтому данную модель использовать не рекомендуется.
6. Увеличение числа соседей с $N = 10$ до 30 не приводит к какому-либо улучшению качества. Для получения хороших результатов стоит использовать $N = 10$ соседей.

Разумеется, работа оставляет открытыми некоторые естественные вопросы. Перечислим те из них, ответы на которые мы собираемся получить, продолжив исследования.

1. Стоит проверить гипотезу о том, может ли метрика EpFound лучше измерять качество ранжирования, чем метрика pFound. Имеется в виду, например, ситуация, изображенная на рис. 4 для модели с весом 0,7 при использовании хостовых признаков. Возможно, различие в поведении pFound и EpFound объясняется плохим поведением самого pFound до добавления 400-го дерева в композицию.
2. Кроме того, мы планируем найти зависимость оптимального веса по критерию гладкости от количества формул для усреднения. Исходя из графика доверительных интервалов можно рассматривать гипотезу о том, что оптимальное значение веса меняется в зависимости от количества формул в усреднении.
3. Наконец, планируется по заданному множеству данных найти критерий выбора веса. Как следует из результатов экспериментов, оптимальное значение веса может быть разным для различных моделей.

Литература

[1] *Burges C. J. C.* From RankNet to LambdaRank to LambdaMART: An overview. Microsoft Research Technical Report, 2010. `http://research.microsoft.com/pubs/132652/msr-tr-2010-82.pdf`.

[2] *Воронцов К. В.* Методы обучения ранжированию. Лекции ШАД. `http://www.machinelearning.ru/wiki/images/8/89/Voron-ML-Ranking-slides.pdf`.

[3] *Гулин А., Карпович П., Расковалов Д., Сегалович И.* Оптимизация алгоритмов ранжирования методами машинного обучения // РОМИП-2009. `http://romip.ru/romip2009/15_yandex.pdf`.

[4] *Wilcoxon F.* Individual comparisons by ranking methods // Biometrics Bull., 1945. Vol. 1. No. 6. P. 80–83. `http://sci2s.ugr.es/keel/pdf/algorithm/articulo/wilcoxon1945.pdf`.

[5] *Buffoni D., Calauzenes C., Gallinari P., Usunier N.* Learning scoring functions with order-preserving losses and standardized supervision // NIPS, 2011. http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Buffoni_447.pdf.

[6] *Calauzenes C., Usunier N., Gallinari P.* On the (non-)existence of convex, calibrated surrogate losses for ranking // NIPS, 2012. http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2012_0118.pdf.

[7] *Shia Y., Karatzogloub A., Baltrunasb L., Larsona M., Hanjalic A.* xCLiMF: Optimizing expected reciprocal rank for data with multiple levels of relevance // Conference on Recommender Systems, 2013.

[8] *Guiver J., Snelson E.* Bayesian inference for Plackett–Luce ranking model // 26th Conference (International) on Machine Learning Proceedings, 2009. http://research.microsoft.com/pubs/81134/plackett.pdf.

Поступила в редакцию 03.09.2016

On a probabilistic model for smoothing discrete ranking quality metrics*

***N. A. Volkov*[1,2] *and M. E. Zhukovskii*[1,2]**

nikitavolkov@yandex-team.ru; zhukmax@yandex-team.ru

[1]Moscow Institute of Physics and Technology
9 Institutskiy Per., Dolgoprudny, Moscow Region, Russia

[2]Yandex, 16 Leo Tolstoy Str., Moscow, Russia

The information retrieval aim is to find relevant documents by given user's query. One of the main information retrieval task is the ranking problem of searching results. Currently, the method of getting relevant function with machine learning methods based on train sample is widely distributed. The quality of ranking is assessing with quality metrics, for example, pFound. However, most of them are discrete, it is difficult for feature selection. The aim of this work is to get a smoothing analogue of the discrete metric. Some different definitions of smoothing metric have been considered, the best of which has been found by the smoothing criteria and the criteria of a similarity to discrete metric by means of experiments. In comparence with smoothing criteria, it is difficult to obtain numerical description of the similarity criteria because there are a lot of different cases of metric behavior. So, the authors decided not to use numerical description. As a result of the large number of experiments, an optimal model of a smoothing metric has been got.

Keywords: *ranking problem; quality metrics; smoothing metrics; discrete metrics; metric pFound*

DOI: 10.21469/22233792.2.4.04

References

[1] Burges, C. J. C. 2010. From RankNet to LambdaRank to LambdaMART: An overview. Microsoft Research Technical Report. Available at: http://research.microsoft.com/pubs/132652/msr-tr-2010-82.pdf (accessed December 29, 2016).

*The research was supported by Yandex.

[2] Vorontsov, K. V. Metody obucheniya ranzhirovaniyu [Learning to rank]. Yandex School of Data Analysis lectures. Available at: http://www.machinelearning.ru/wiki/images/8/89/Voron-ML-Ranking-slides.pdf (accessed December 29, 2016).

[3] Gulin, A., P. Karpovich, D. Raskovalov, and I. Segalovich. 2009. Optimizatsiya algoritmov ranzhirovaniya metodami mashinnogo obucheniya [Optimizing the ranking algorithms by machine learning methods]. *ROMIP-2009*. Available at: http://romip.ru/romip2009/15_yandex.pdf (accessed December 29, 2016).

[4] Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bull.* 1(6):80–83. Available at: http://sci2s.ugr.es/keel/pdf/algorithm/articulo/wilcoxon1945.pdf (accessed December 29, 2016).

[5] Buffoni, D., C. Calauzenes, P. Gallinari, and N. Usunier. 2011. Learning scoring functions with order-preserving losses and standardized supervision. *NIPS*. Available at: http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Buffoni_447.pdf (accessed December 29, 2016).

[6] Calauzenes, C., N. Usunier, and P. Gallinari. 2012. On the (non-)existence of convex, calibrated surrogate losses for ranking. *NIPS*. Available at: http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2012_0118.pdf (accessed December 29, 2016).

[7] Shia, Y., A. Karatzogloub, L. Baltrunasb, M. Larsona, and A. Hanjalic. 2013. xCLiMF: Optimizing expected reciprocal rank for data with multiple levels of relevance. *Conference on Recommender Systems*.

[8] Guiver, J., and E. Snelson. 2009. Bayesian inference for Plackett–Luce ranking model. *26th Conference (International) on Machine Learning Proceedings*. Available at: http://research.microsoft.com/pubs/81134/plackett.pdf (accessed December 29, 2016).

Received September 3, 2016

Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления*

Н. Д. Смелик, А. А. Фильченков
smelik@rain.ifmo.ru; afilchenkov@corp.ifmo.ru
Университет ИТМО, Россия, г. Санкт-Петербург, Кронверкский проспект, 49

Целью данной работы является создание мультимодальной тематической модели для изображений и текстов. Предложен подход к построению такой модели на основе векторного представления текстов и изображений. Векторы значимых слов строятся за счет применения Word2Vec, для изображений — как выход последнего неполносвязного слоя сверточной нейронной сети. Предложены алгоритм обучения тематической модели по коллекции аннотированных изображений, а также алгоритмы аннотирования нового изображения и иллюстрирования нового текста. Эксперименты показали, что предложенная модель превосходит аналоги в задаче аннотирования изображений и иллюстрирования текстов.

Ключевые слова: *вероятностная тематическая модель; аннотирование изображение; иллюстрирование текста; сверточные нейронные сети; векторное представление слов*

DOI: 10.21469/22233792.2.4.05

1 Введение

Тематическое моделирование — активно развивающаяся с конца 1990-х гг. область машинного обучения, применимая к анализу текстов. Ее появление обусловлено необходимостью обрабатывать огромные объемы цифровых данных, которые стали доступны человеку после появления Интернета. Тематическая модель определяет, к каким темам относится каждый документ, а также из каких терминов состоит каждая тема. Это позволяет эффективно решать задачи тематического поиска [1], категоризации и классификации текстовых документов [2], а также аннотации разного вида данных. Перечисленные задачи находят применение в различных областях, где приходится иметь дело с поиском различной информации схожей тематики.

Изначально тематические модели учитывали при построении только слова, из которых состоят документы, однако в последнее время стали появляться модели, позволяющие учитывать также сопутствующую документу информацию, такую как теги, авторы [3], атрибуты [4] и др. К этому списку следует также отнести тематическую модель для текстов и изображений. Такая модель позволяет автоматически выделять темы из изображений на основе их описания и в дальнейшем использовать ключевые слова тем для описания новых изображений. Это может найти применение в таких задачах, как аннотирование изображений или поиск тематических иллюстраций для текста.

В последнее время появились инструменты, показавшие отличные результаты в области распознавания образов и обработки естественного языка. Это технология глубокого обучения, успешно применяющаяся для распознавания изображений, и модель векторного

*Работа выполнена при финансовой поддержке Правительства Российской Федерации, грант 074-U01.

представления слов, позволяющая учитывать контекст слова. Использование этих технологий может повысить качество построения тематических моделей для текстов и изображений.

Целью данной работы является повышение качества аннотирования изображений и иллюстрирования текста путем разработки тематической модели на основе совместного использования тематического моделирования, глубоких нейронных сетей и векторного представления слов.

В разд. 2 представлены основные понятия тематических моделей и описание исследований, посвященных этой тематике. В разд. 3 кратко описана идея использовать сверточные нейронные сети и векторные представления слов для построения мультимодальной тематической модели, а также описаны соответствующие технологии. В разд. 4 и 5 описаны алгоритмы соответственно обучения модели и ее применения для аннотирования изображений и иллюстрирования описаний. В разд. 6 приведены подробности реализации модели. В разд. 7 описаны вычислительные эксперименты, проведенные для сравнения модели с аналогами, а в разд. 8 — результаты экспериментов. Раздел 9 является заключительным.

2 Тематические модели

В данном разделе приведены необходимые понятия и обозначения, а также краткий обзор исследований в области тематического моделирования. Вначале будет дано определение тематической модели, а также будут сделаны основные предположения. Затем будет дано краткое описание существующих на данный момент тематических моделей.

2.1 Вероятностная модель коллекции документов

Вероятностная тематическая модель представляет собой модель коллекции текстовых документов, которая определяет темы для каждого документа [5]. Интуитивно можно высказать предположение, что документ, принадлежащий какой-либо конкретной теме, будет содержать много специфических для этой темы терминов. Тематическая модель строит на этом предположении математическую модель, которая позволяет, основываясь на статистических данных, предположить темы, к которым относится документ, а также соотношение этих тем в документе.

Основные предположения базовых вероятностных тематических моделей:

1) не важен порядок документов в корпусе;
2) не важен порядок слов в документе, документ — «мешок слов» [6];
3) слова, встречающиеся в большинстве документов (слова общей лексики), не важны для определения тематики документа;
4) слово в разных формах считается одним и тем же словом;
5) коллекцию документов можно рассматривать как случайную, однородную и независимую выборку пар документ–слово (d, w), $d \in D$, $w \in W_d$;
6) каждая тема $t \in T$ описывается неизвестным распределением $p(w|t)$ на множестве слов W, $w \in W$;
7) каждый документ описывается неизвестным распределением $p(t|d)$ на множестве тем T, $t \in T$;
8) гипотеза условной независимости: $p(w|t, d) = p(w|t)$;
9) $p(w|d) = \sum_{t \in T} p(t|d)p(w|t)$ — вероятностная модель порождения данных.

Построить тематическую модель — значит найти матрицы $\Phi = \|p(w|t)\|$ и $\Theta = \|p(t|d)\|$ для коллекции D. Для оценки параметров Φ и Θ используется принцип максимума правдоподобия:

$$p(D;\Phi,\Theta) = C\prod_{d\in D}\prod_{w\in d} p(d,w)^{n_d w} = \prod_{dinD}\prod_{w\in d} p(w|d)^{n_{dw}}\mathrm{Cp}\,(d)^{n_{dw}} \to \max_{\Phi,\Theta}.$$

где $n_d w$ — число вхождений термина w в документ d; $\mathrm{Cp}\,(d)^{n_{dw}}$ — константа; C — нормировочный множитель, зависящий только от чисел n_{dw}.

Чаще для удобства используется логарифм максимума правдоподобия:

$$L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta};$$

$$\sum_{w\in W}\varphi_{wt} = 1;\varphi \geqslant 0; \quad \sum_{t\in T}\theta_{td} = 1;\theta_{td} \geqslant 0.$$

2.2 Обзор существующих тематических моделей

Тематическое моделирование берет свое начало из работы Пападимитриу, Томаки, Рагавана и Вемполы в 1998 г. [7]. В этой работе для построения тематической модели для коллекции текстовых документов использовался метод латентно-семантического анализа (Latent Semantic Analysis, LSA), который был разработан и запатентован в 1988 г. группой ученых, возглавляемых Скотом Дирверстером [8]. Определение слов, из которых состоят темы, а также определение тем, к которым принадлежит документ, осуществляется с помощью применения к матрице «Слова-на-Документы» *сингулярного матричного разложения* (Singular Value Decomposition, SVD) [9]. С помощью этого метода исходная матрица A представлялась в виде произведения трех матриц $A = USV^{\intercal}$, где матрицы U и V — ортогональные, а матрица S — диагональная, содержащая на главной диагонали числа, называемые сингулярными. Основная идея LSA заключалась в том, что матрица A', содержащая только первые k линейно независимые компоненты A, содержит основную структуру зависимостей, которые присутствуют в исходной матрице «Слова-на-Документы». В результате каждое слово и документ представляются в виде вектора в общем пространстве размерности k, и близость между комбинациями слов и/или документов вычисляется с помощью скалярного произведения векторов.

Основным недостатком этой модели является сильное уменьшение скорости вычислений с увеличением объема входных данных [8]. Также стоит отметить, что у SVD, как и у любого другого матричного разложения, отсутствует явное лингвистическое обоснование, поэтому оценка работы модели и интерпретация ее результатов не всегда понятны.

Как дальнейшее развитие LSA Томас Хоффман предложил в 1999 г. новую модель под названием *вероятностный латентно-семантический* анализ (Probabilistic Latent Semantic Analysis, PLSA) [10]. Данная модель основывается на статистической модели скрытых классов и может быть представлена как вероятностная тематическая модель, описанная в предыдущем пункте.

В PLSA для максимизации логарифма правдоподобия используется EM (expectation-maximization) алгоритм [11]. На E-шаге для текущих значений параметров φ_{wt} и θ_{td} с помощью формулы Байеса вычисляются условные вероятности $p(t|d,w)$ для всех тем $t \in T$ для каждого термина $w \in d$ в каждом документе d:

$$H_{dwt} = p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_{s\in T}\varphi_{ws}\theta_{sd}}.$$

На M-шаге, наоборот, по условным вероятностям тем H_{dwt} вычисляется новое приближение параметров φ_{wt} и θ_{td}.

Преимуществом PLSA по сравнению с предыдущей моделью LSA является то, что PLSA основан на статистике и поэтому лучше подходит для применения на практике. Основным недостатком данного подхода считается невозможность управлять разреженностью матриц Φ и Θ. Также стоит отметить, что число параметров PLSA растет линейно с числом документов в коллекции, что приводит к переобучению модели.

С целью преодоления этих недостатков в 2003 г. Девидом Блеем, Эндрю Ыном и Майклом Джорданом была разработана модель, которая является логическим продолжением PLSA — *латентное размещение Дирихле* (Latent Dirichlet Allocation, LDA) [12]. На данный момент эта модель является самой популярной в тематическом моделировании, и на ее основе было создано большое число других моделей для разного рода задач.

Модель LDA основана на том же принципе максимизации логарифма правдоподобия что и PLSA, однако дополнительно было выдвинуто предположение, что векторы документов $\theta_d = (\theta_{td}) \in \mathbb{R}^{|t|}$ и векторы слов $\varphi_t = (\varphi_{wt}) \in \mathbb{R}^{|w|}$ порождаются распределением Дирихле с параметрами $\alpha \in \mathbb{R}^{|t|}$ и $\beta \in \mathbb{R}^{|w|}$ соответственно. Это позволяет управлять разреженностью матриц Φ и Θ, что приводит к получению более корректного набора тем по сравнению с PLSA.

Основным недостатком LDA является слабое лингвистическое обоснования использования распределения Дирихле в качестве порождающего распределения для θ_d и φ_t. На самом деле это предположение кажется весьма произвольным. Распределение Дирихле было выбрано для удобства вычислений, и вовсе необязательно соответствует реальной порождающей модели.

Последняя модель для определения тем текстовых документов была предложена сравнительно недавно и является альтернативой байесовскому подходу и графическим моделям, которые использовались в PLSA и LDA. Данная модель называется «Аддитивная регуляризация тематических моделей» (Additive Regularization of Topic Models, ARTM) и была предложена в 2014 г. Константином Воронцовым [13].

В основу создания модели легло решение проблемы неустойчивости и неединственности матричного разложения $\Phi\Theta$. Общий подход преодоления этой проблемы состоит в использовании регуляризации. Он заключается в введении ограничений на Φ и Θ, что в конечном итоге приводит к сужению пространства решений.

Соответственно подход ARTM основан на идее многокритериальной регуляризации. Он позволяет строить модели, удовлетворяющие многим ограничениям одновременно. Каждое ограничение формализуется в виде регуляризатора — оптимизационного критерия $R_i(\Phi, \Theta) \to \max$, зависящего от параметров модели. Взвешенная сумма таких критериев $R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi\Theta)$ максимизируется совместно с логарифмом правдоподобия

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

при тех же ограничениях нормировки и неотрицательности.

2.3 Мультимодальные тематические модели текстов и изображений

Тематическая модель, которая строится не только по словам, но и по любым терминам другой модальности, называется *мультимодальной*. К таким терминам могут относиться, например, авторство текста, время его написания или число людей, выразивших в социальной сети одобрение этому тексту. Для разных модальностей вероятностные распределения над терминами должны строиться раздельно. В качестве расширения ARTM в 2015 г. была предложена модель, являющаяся мультимодельным обобщением ARTM [14] с открытой

библиотекой BigARTM[1] [**?**]. Однако поскольку пространство изображений вычислительно неперечислимо, то любое построенное над ним вероятностное распределение будет характеризовать исключительно коллекцию, по которой оно было построено. В связи с этим применить мультимодальную ARTM к данной задаче напрямую невозможно.

Идея построения мультимодальных тематических моделей для текстов и изображений не нова. Первые работы в этом направлении были проделаны Девидом Блеем и Майклом Джорданом в работе [16], в которой была предложена модель на основе LDA для автоматической аннотации изображений под названием Correspondence LDA (CorrLDA). В данной модели изображения сначала сегментируются с помощью алгоритма N-cuts [17], затем для каждого сегмента извлекается вектор признаков, которыми выступают размер, позиция, цвет, текстура и форма, представленные в виде действительных значений. После этого формируются пары (вектор признаков изображения, набор слов описания). Вектор признаков изображения предполагается порожденным многофакторным гауссовым распределением с диагональной матрицей ковариации, а набор слов описания предполагается порожденным полиномиальным распределением на словаре. После этого модель можно рассматривать с точки зрения порождающего процесса, который сначала генерирует дескрипторы изображения, а затем генерирует подпись в виде слова. В частности, сначала создаются N дескрипторов с помощью модели LDA, а затем для каждого из M слов описания выбирается одна из областей изображения, и после этого слово привязывается к выбранному региону.

В работах [18, 19] были предложены модели MixLDA и sLDA, основным отличием от CorrLDA в которых стало использование алгоритма SIFT [20] для извлечения признаков из изображений. Признаки, извлеченные из изображений в ходе обучения моделей, кластеризовались с помощью алгоритма k-means, что позволяло представить изображения в виде «мешка визуальных слов» и объединить их с текстовой модальностью. Для извлечения тем в MixLDA используется вариант LDA, описанный в [12]. В sLDA используется усовершенствованная модель, описанная в [21].

3 Предлагаемый подход и описание технологий

В данной работе предлагается использовать подход к построению тематических моделей текстов и изображений, совмещающий в себе технологии, подробно описываемые в этой главе, а именно: глубокие нейронные сети и векторные представления слов. Предполагается, что для обучения на вход модели подается коллекция изображений, сопровождаемых описаниями. Для извлечения признаков изображения предлагается использовать сверточную нейронную сеть. Каждое значимое слово в аннотации предлагается заменить на его векторное представление, это позволит учитывать также контекст слова. Этот шаг позволит расширить словарь, так как будут учитываться также слова, употребляющиеся в схожих контекстах. Затем набор признаков изображения представляется в виде псевдодокумента, в котором словами будут векторные представления слов из описания к изображению. Нахождение скрытых тем в таких псевдодокументах позволит использовать полученную модель в задачах аннотирования изображений и иллюстрирования текста. Ожидается, что полученная модель повысит качество аннотирования изображения и иллюстрирования текста.

[1] `http://bigartm.org`.

3.1 Сверточные нейронные сети

Нейронные сети — мощная математическая модель, основанная на знаниях, полученных при изучении нейронных связей в мозге. Она представляет собой набор связанных и взаимодействующих между собой искусственных нейронов. Такие нейроны обычно очень просты, но, будучи соединенными в сеть, они способны решать довольно сложные задачи.

Одной из задач, решаемых нейронными сетями, является задача компьютерного зрения (computer vision, CV) [22], а именно: задача распознавания изображений. Наибольших успехов в решении этой задачи добились *сверточные нейронные сети* (Convolutional Neural Network, CNN) [23]. Сверточная нейронная сеть — глубокая нейронная сеть, для которой сделано явное предположение, что на вход подается изображение. Это позволяет добавить в архитектуру такой сети определенные свойства, которые улучшают эффективность и снижают число параметров по сравнению с обычными сетями.

Сверточная сеть использует три типа слоев: сверточные (convolutional), субдискретизирующие (max-pooling) и полносвязные (full connected). Чередование первых двух типов слоев позволяет получить из изображения набор карт признаков (feature map), полносвязный слой применяется для классификации полученных наборов.

Каждый нейрон сверточного слоя отвечает за применение операции свертки части изображения, поданного на вход нейрону, с фильтром, реагирующим на простые линии. Часть изображения, к которой применяется операция свертки, называется окном. За счет смещения этого окна нейрон обрабатывает все изображение. В результате применения фильтра к каждой позиции изображения на выходе сверточного нейрона получается карта признаков. При этом размер этой карты может быть как больше исходного изображения, так и меньше него. Это зависит от способа смещения окна фильтра.

Субдискретизирующие слои обычно находятся между сверточными. Основное назначение субдискретизирующего слоя заключается в уменьшении размерности представления, полученного на предыдущем сверточном слое. Субдискретизирующий нейрон также использует окно для обхода изображения. Для набора значений, попадающих в окно, применяется некоторая функция (например, функция, выбирающая максимум из значений в пределах окна), значение которой записывается на выход нейрона.

Сверточные нейронные сети обладают замечательной особенностью: если убрать полносвязные слои, которые используются для классификации изображений, получим набор признаков, которые характеризуют наше изображение. Также отбрасывание полносвязных слоев позволяет абстрагироваться от категорий изображений, используемых при обучении такой сети, и извлекать признаки из любых изображений. В работе [24] была экспериментально показана успешность применения такой техники извлечения признаков.

3.2 Векторное представление слов

Векторное представление слова (word embedding) [25] — параметризованное отображение слова на пространство большой размерности $\mathbb{R}^n$. Такое представление позволяет выразить семантическое и синтаксическое значение слова в виде вектора фиксированной длины.

В последнее время для порождения векторных представлений наибольшей популярностью пользуется набор алгоритмов, разработанный Томасом Миколовым под названием Word2Vec [26]. В основе Word2Vec лежит нейронная сеть, которая обучается на большом текстовом корпусе и на выходе генерирует для каждого слова из словаря его векторное представление большой размерности. Основной задачей Word2Vec является максимизация

расстояния между векторами слов, близких по смыслу, и минимизация расстояний между векторами слов, различных по смыслу [27].

В Word2Vec используются две модели для векторного представления слов: skip-gram и continuous bag of word (CBOW). Отличие этих моделей заключается в том, что модель skip-gram предсказывает контекст по слову, т. е. данная модель умеет по входному слову w_i предсказывать $(w_{i-3}, w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}, w_{i+3})$ — набор слов, которые употребляются чаще всего вместе со словом, поданным на вход. Модель CBOW, наоборот, предсказывает слово по данному контексту, т. е. решает задачу, обратную задаче, которую решает skip-gram. Этот метод подходит наилучшим образом для создания вектора описания слов, так как на выходе выдает вектор контекста, в котором слово часто используется, что позволит учитывать этот контекст при построении тематической модели.

Одним из подходов к настройке параметров модели skip-gram является максимизация логарифма условной вероятности использования слова в контексте. Эту вероятность можно выразить в виде softmax функции:

$$(c|w;\Theta) = \frac{\exp(v_c v_w)}{\sum_{c^f \in C} \exp(v_c v_w)},$$

где v_w — вектор, представляющий слово w; v_c — вектор, который представляет контекст слова w; C — множество всех доступных контекстов. Однако так как контекстов может быть довольно много, то их полный перебор является довольно длительной задачей.

Для уменьшения количества вычислений в word2vec используется метод, разработанный Томасом Миколовым под названием *отрицательное сэмплирование* (negative sampling). В данном методе модель обучается на парах (w, c), где w — слово, а c — контекст этого слова. Вероятность $p(D = 1|w, c)$ означает вероятность того, что пара (w, c) пришла из данных для обучения. Вероятность $p(D = 0|w, c)$ означает вероятность того, что пара (w, c) не является парой из данных для обучения. Тогда задача метода negative sampling сводится к максимизации вероятности того, что пара (w, c) является парой из данных для обучения:

$$\arg\max_{\Theta} \sum_{(w,c)\in D} \log p(D = 1|w, c; \Theta),$$

где

$$p(D = 1|w, c; \Theta) = \frac{1}{1 + \exp(-v_c v_w)}.$$

Эта задача имеет тривиальное решение, если установить $p(D = 1|w, c; \Theta) = 1$ для всех пар (w, c). Однако тогда все векторы получат одно и то же значение, что крайне не желательно.

Один из способов усовершенствовать тривиальное решение — запрещать некоторые комбинации пар (w, c). Это достигается путем генерации множества D' — случайных некорректных пар (w, c), которые не принадлежат данным для обучения (название метода negative sampling происходит из использования множества случайно выбранных неправильных (negative) пар (w, c)). В итоге задача приобретает следующий вид:

$$\arg\max_{\Theta} \sum_{(w,c)\in D} \log p(D = 1|w, c; \Theta) + \sum_{(w,c)\in D'} \log p(D = 0|w, c; \Theta).$$

4 Алгоритм построения модели

Модель для обучения получает на вход корпус, состоящий из изображения и подходящего к нему описания. Описание может быть различной длины, а также их может быть несколько, относящихся к одному изображению.

После обучения модель возвращает матрицы Φ, описывающую распределение векторов слов на темы, и Θ, описывающую распределение тем на образы изображений.

4.1 Предварительная обработка данных

Так как на вход модели поступают сырые данные в виде набора пар изображение и его описание, необходимо подготовить эти данные для дальнейшей обработки. Основная задача предварительной обработки заключается в обработке описаний изображений.

Каждое описание разбивается на слова, при этом нет смысла различать разные формы одного и того же слова, так как это приведет к разрастанию словаря и ухудшению качества модели. Для приведения слов к нормальной форме используются лемматизация или стемминг.

Лемматизация — процесс приведения слова к нормальной форме. Применительно к русскому языку процесс лемматизации приводит существительные в любой форме к форме именительного падежа единственного числа, прилагательные в любой форме — к форме именительного падежа единственного числа мужского рода, а глаголы, причастия и деепричастия в любой форме — к форме глагола в инфинитиве. Лемматизация требует больших временны́х затрат, а также заставляет хранить большие словари. Это оправдано для языков, относящихся к агглютинативным или синтетическим.

На практике чаще всего применяется стемминг. *Стемминг* — это процесс нахождения основы слова. Самый простой способ стемминга заключается в отбрасывании окончания, также существуют и другие, более сложные варианты стемминга [28–31].

После процесса лемматизации или стемминга отбрасываются *стоп-слова* — слова, встречающиеся почти во всех документах. Этот шаг мотивирован тем, что такие слова бесполезны для тематического моделирования, так как они, скорее всего, будут встречаться во всех темах, но на самом деле они не будут характеризовать тему. Это так называемые слова общей лексики. К ним относятся предлоги, союзы, местоимения, числительные, прилагательные, некоторые глаголы и наречия. Число таких слов обычно варьируется в пределах нескольких сотен, так что для их идентифицирования заранее составлен словарь стоп-слов. Заметим, что отбрасывание стоп-слов практически не влияет на длину словаря.

В конечном итоге формируется словарь, состоящий из всех слов, которые встречались во всех описаниях и не были отброшены на предыдущем шаге. При этом каждому слову присваивается идентификатор — номер, под которым он записан в словаре.

Для оценки веса каждого слова в описании используется неотрицательная мера TF-IDF (term frequency — inverse document frequency). Она описывается следующим выражением:

$$\text{tfidf}\,(t, d, D) = \text{tf}\,(t, d) \cdot \text{idf}\,(t, D).$$

Здесь

$$\text{tf}\,(t, d) = \frac{|t|}{\sum_{k \in d} |t_k|}; \quad \text{idf}\,(t, D) = \log \frac{|D|}{|d_i, t|},$$

где $|t|$ — число вхождений термина t в документ; $\sum_{k \in d} |t_k|$ — общее число слов в документе; $|D|$ — число документов в корпусе; $|d_i, t|$ — число документов в коллекции, в которых встречается термин t.

Вес каждого термина также добавляется в словарь.

В итоге после предварительной обработки для каждого изображения формируют взвешенный набор слов, которые наиболее важны при описании изображения:

$$I \leftrightarrow ((\sigma_1, w_1), (\sigma_2, w_2) \ldots, (\sigma_n, w_n)),$$

где σ_i — вес слова w_i.

4.2 Алгоритм обучения модели

Процесс обучения начинается с обучения моделей для векторного представления слов и векторного представления изображений. Для обучения модели векторного представления слов используется модель skip-gram, для обучении векторного представления изображений — CNN, описанные в предыдущем разделе.

Для построения тематических моделей текстовых документов необходимо сформировать матрицу вхождения каждого слова в каждый документ, т. е. набор векторов $w_j = p_1, p_2, \ldots, p_n$, $j = 1, \ldots, m$, где w_j — j-е слово из словаря; p_i — вероятность появления слова w_j в i-м документе; n — число документов в коллекции; m — число слов в словаре.

В рассматриваемой задаче вектор признаков изображения $\mathbf{i}$ будет представлен в виде псевдодокумента, в котором словами будут векторные представления слов $\mathbf{w}$ — из аннотации к изображению. Это позволит построить модель, которая сможет получать распределения тем для каждого изображения, основываясь на словах из его описания. В качестве вероятности появления слова в псевдодокументе будем использовать меру TF-IDF, описанную ранее.

Полученная матрица будет иметь вид:

$$F = (p_{\mathbf{wi}})_{|W| \times |I|}, \quad p_{\mathbf{weie}} = \text{tfidf}\,(\mathbf{w}, \mathbf{i}, I),$$

где $\mathbf{w}$ — векторное представление слова w из словаря W, $\mathbf{i}$ — векторное представление изображения из множества псевдодокументов I.

Для выделения тем будем использовать подход, аналогичный тематической модели ARTM, описание которой дается в разд. 2. Будем решать задачу приближенного представления матрицы F в виде произведения $F \approx \Phi\Theta$ двух матриц. Для данной задачи это будут матрица векторных представлений слов на темы Φ и матрица тем на векторы признаков изображений Θ.

Как уже было сказано в разд. 2, данная задача решается путем максимизации логарифма максимума при ограничениях нормированности и неотрицательности столбцов матриц Φ и Θ:

$$L(\Phi, \Theta) = \sum_{\mathbf{i} \in I} \sum_{\mathbf{w} \in i} p_{\mathbf{wi}} \ln \sum_{t \in T} \varphi_{\mathbf{w}t} \theta_{t\mathbf{i}} \to \max_{\Phi, \Theta},$$

где

$$\sum_{\mathbf{w} \in W} \varphi_{\mathbf{w}t} = 1; \varphi_{\mathbf{w}t} \geqslant 0; \quad \sum_{t \in T} \theta_{t\mathbf{i}} = 1; \theta_{t\mathbf{i}} \geqslant 0\,.$$

Для решения этой задачи применим EM-алгоритм. Перед первой итерацией задаем начальные приближения для параметров $\varphi_{\mathbf{w}t}$ и $\theta_{t\mathbf{i}}$.

Затем на E-шаге по текущим значениям параметров $\varphi_{\mathbf{w}t}$ и $\theta_{t\mathbf{i}}$ с помощью формулы Байеса вычисляются условные вероятности $p(t|\mathbf{i}, \mathbf{w})$ для всех тем $t \in T$ для каждого векторного представления слова $\mathbf{w} \in \mathbf{i}$ в каждом векторе признаков изображений $\mathbf{i}$:

$$H_{iewet} = p(t|\mathbf{i}, \mathbf{w}) = \frac{p(\mathbf{w}|t)p(t|\mathbf{i})}{p(\mathbf{w}|\mathbf{i})} = \frac{\varphi_{\mathbf{w}t}\theta_{t\mathbf{i}}}{\sum_{s \in T} \varphi_{\mathbf{w}s}\theta_{s\mathbf{i}}}.$$

На M-шаге по полученным условным вероятностям тем $H_{\mathbf{i}\mathbf{w}t}$ вычисляется новое приближение параметров $\varphi_{\mathbf{w}t}$ и $\theta_{t\mathbf{i}}$. Псевдокод данного EM-алгоритма приведен на листинге 1.

Алгоритм 1 EM-алгоритм обучения

Вход: коллекция изображений с описаниями I, число тем $|T|$, начальные приближения Φ и Θ
Выход: распределения Φ и Θ
повторять
обнулить $n_{\mathbf{w}t}, n_{\mathbf{i}t}, n_t$ для всех $\mathbf{i} \in I$, $\mathbf{w} \in W$, $t \in T$;
для всех $\mathbf{i} \in I$, $\mathbf{w} \in ie$
$Z = \sum_{t \in T} \varphi_{\mathbf{w}t}\theta_{t\mathbf{i}}$
для всех $t \in T$, что $\varphi_{\mathbf{w}t}\theta_{t\mathbf{i}} > 0$
увеличить $n_{\mathbf{w}t}, n_{t\mathbf{i}}, n_t$ на $\delta = n_{\mathbf{i}w}\varphi_{\mathbf{w}i}\theta_{ti}/Z$
пока Φ и Θ не сойдутся.

Однако, как показано в работе [32], искомое стохастическое матричное разложение $\Phi\Theta$ определено не единственным образом, а с точностью до невырожденного преобразования $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$ при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические, и задача тематического моделирования в общем случае имеет бесконечно много решений. Это ведет к неустойчивости EM-алгоритма. Для решения этой проблемы предлагается ввести дополнительные ограничения $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$ (регуляризаторы) на Φ и Θ для того, чтобы сузить множество решений.

За счет этого задача сводится к максимизации линейной комбинации критериев L и R_i с неотрицательными коэффициентами регуляризации τ_i при все тех же ограничениях неотрицательности и нормировки матриц Φ и Θ:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta); \quad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}.$$

В итоге для решения задачи матричного разложения будем применять модифицированный EM-алгоритм, приведенный на листинге 2.

Алгоритм 2 EM-алгоритм для модели ARTM

Вход: коллекция изображений с описаниями I, число тем $|T|$, начальные приближения Φ и Θ
Выход: распределения Φ и Θ
повторять
обнулить $n_{\mathbf{w}t}, n_{\mathbf{i}t}, n_t$ для всех $\mathbf{i} \in I$, $\mathbf{w} \in W$, $t \in T$;
для всех $\mathbf{i} \in I$, $\mathbf{w} \in ie$
$Z = \sum_{t \in T} \varphi_{\mathbf{w}t}\theta_{t\mathbf{i}}$
для всех $t \in T$, что $\varphi_{\mathbf{w}t}\theta_{t\mathbf{i}} > 0$
увеличить $n_{\mathbf{w}t}, n_{t\mathbf{i}}, n_t$ на $\delta = n_{\mathbf{i}w}\varphi_{\mathbf{w}i}\theta_{ti}/Z$
$\varphi_{\mathbf{w}t} \propto (n_{\mathbf{w}t} + \varphi_{\mathbf{w}t}\partial R/\partial\varphi_{\mathbf{w}t})_+$ для всех $\mathbf{w} \in W$, $t \in T$
$\theta_{t\mathbf{i}} \propto (n_{\mathbf{i}t} + \theta_{t\mathbf{i}}\partial R/\partial\theta_{t\mathbf{i}})_+$ для всех $\mathbf{i} \in I$, $t \in T$
пока Φ и Θ не сойдутся.

В результате работы данного алгоритма будут получены матрицы Φ и Θ, которые выражают условные распределения на множестве векторных представлений слов для каждой темы и условные распределения на множестве тем для каждого вектора образов изображения. Это и есть искомая тематическая модель.

5 Алгоритмы применения модели

Обученная модель получает на вход либо только текст, по которому требуется подобрать иллюстрации, либо только изображение, к которому требуется подобрать описание.

5.1 Алгоритм генерации аннотаций по изображению

Имея разложение «матрицы векторные представления слов на векторы образов изображений» на матрицы Φ и Θ, можно относительно легко генерировать описания изображений. Последовательность шагов для нахождения слов, подходящих к описанию изображения, приведена на рис. 1. Опишем подробнее каждый шаг.

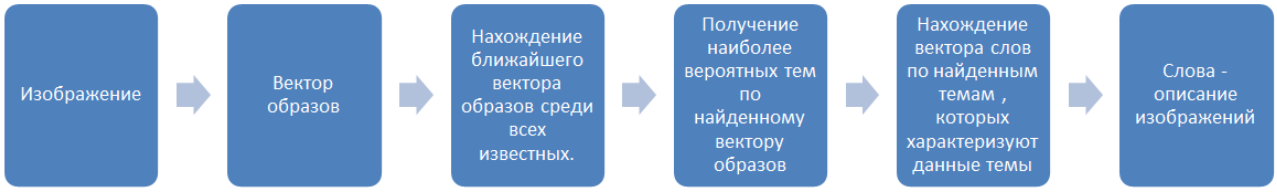


Рис. 1 Схема алгоритма генерации аннотаций по изображению

На вход модели подается изображение. Так как изображение может быть произвольное, а CNN принимает изображения определенного размера, необходимо изменить размер изображения. После этого изображение подается на вход CNN, которая выдает вектор признаков данного изображения. Далее модель начинает поиск вектора, ближайшего к тому, который был подан на вход, среди всех векторов, известных модели. Для найденного вектора из матрицы Θ извлекается распределение тем, полученное во время создания модели.

В результате по найденным темам с помощью матрицы Φ извлекаются векторы слов, которые наилучшим образом характеризуют контекст данной темы. С помощью этих векторов можно найти слова, которые чаще всего употребляются в этом контексте. Это и будет описание поданного на вход модели изображения.

5.2 Алгоритм поиска изображений по текстам

Также имея разложения матрицы F на матрицы Φ и Θ, можно реализовать алгоритм поиска изображений по тексту, или задачу иллюстрации текста.

Последовательность шагов, необходимых для нахождения изображений по текстовому описанию, приведена на рис. 2. Рассмотрим каждый шаг подробнее.

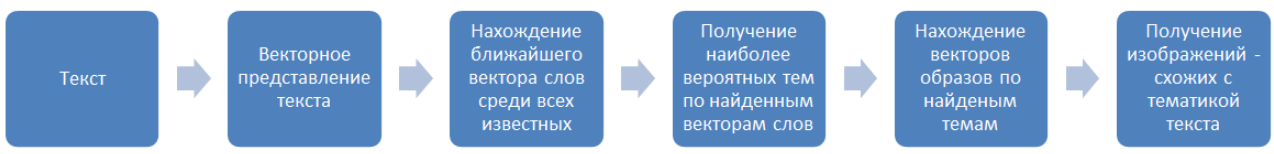


Рис. 2 Схема алгоритма поиска изображений по текстам

На вход модели подается текст. Текст проходит предобработку, описанную в п. 4.1, после чего для каждого термина в тексте находится его векторное представление. Далее текст представляется в виде нового документа, для которого нужно найти распределение тем, т. е. получить вектор θ распределений тем для нового документа. Это можно сделать с помощью алгоритма для обучения модели с той разницей, что матрица Φ уже известна.

После получения распределения тем для текстового документа для каждой темы, вероятность которой в документе не равна 0, находим изображение, максимально соответствующее данной теме. Этот набор изображений и будет иллюстрациями к тексту.

6 Особенности реализации обучения модели

Вся реализация была написана на языке Python версии 2.7. Здесь и далее все используемые библиотеки являются библиотеками для языка Python.

Для получения векторов признаков изображений была использована предобученная CNN под названием VGG-16 [33]. Эта сверточная нейронная сеть показала хорошие результаты в различных соревнованиях по классификации изображений. У этой CNN убирается последний softmax слой, в результате чего сеть выдает 4096-мерный вектор признаков. Использование предобученной сети обусловливалось тем, что процесс обучения такой сложной сети требует большого количества времени и вычислительных ресурсов, при этом качество полученной модели могло получиться хуже. Для взаимодействия с нейронной сетью была использована библиотека Keras[2] для построения нейронных сетей.

Для получения векторных представлений слов был использован также предобученый набор моделей под названием Word2Vec, разработанный кампанией Google [26]. В качестве ее реализации использовалась реализация, предоставленная библиотекой Gensim[3]. Так как для обучения Word2Vec так же, как и для CNN, требуется большое количество времени и вычислительных ресурсов, а также корпус порядка нескольких миллионов слов, было решено также использовать предобученную модель. В качестве предобученной модели для Word2Vec использовалась модель skip-gram с окном размера 10, построенная на основе корпуса Wikipedia[4]. На выходе модель отдает 1000-мерный вектор.

В качестве тематической модели было решено применять ARTM, описанную в разд. 2. В качестве ее реализации использовалась библиотека BigARTM. Выбор данной модели основывается на применении в ней регуляризаторов, использование которых позволяет сравнивать эту модель с моделями PLSA и LDA.

Для предобработки входных данных был написан модуль, осуществляющий предобработку входных данных, генерирующий необходимые словари, а также создающий специальные пакеты (batch), необходимые для библиотеки BigARTM. Описания для каждого изображения были объединены в одно, и ему присваивался уникальный идентификатор; этот же идентификатор присваивался изображению, к которому относились описания. Из описаний изображений были удалены все знаки препинания, а также все неалфавитные символы. После этого описание разбивалось на набор слов, в которые входили только существительные, глаголы и прилагательные. Также в словарь не входили некоторые слова общей лексики, которые могли бы помешать выделить темы. Для удаления таких слов был составлен словарь из 600 стоп-слов, не несущих никакой полезной нагрузки для определения темы. Из оставшихся слов был составлен словарь, в котором каждому слову был

[2] `http://keras.io/`.

[3] `https://github.com/piskvorky/gensim/`.

[4] `https://github.com/idio/wiki2vec`.

присвоен уникальный идентификатор, а в каждом описании слова были заменены на соответствующий идентификатор. При дальнейшем построении тематической модели применялись только идентификаторы слов. После этого каждое слово в словаре было заменено на соответствующее векторное представление, и словарь сохранялся. Для каждого идентификатора в описании производился подсчет меры TF-IDF. После этого формировался batch файл.

Каждое изображение сжималось до размеров 224×224, так как CNN обрабатывает изображения именно такого размера. После этого из изображения извлекался вектор признаков, которому присваивался идентификатор, полученный при обработке описаний этого изображения. Далее формировался и сохранялся словарь. При дальнейшем построении тематической модели использовались только идентификаторы признаков изображений.

После предобработки получалось три файла: batch-файл, представляющий входные данные для BigARTM; словарь соответствия идентификатора слова и его векторного представления; словарь соответствия идентификатора изображения и его векторного представления.

В конечном итоге после обработки batch-файла библиотекой BigARTM сохранялись матрицы Φ и Θ, полученные в результате построения тематической модели.

7 Вычислительные эксперименты

С помощью реализованной модели были проведены эксперименты по поиску изображений по тексту и текстов по описанию.

Для тестирования был использован набор данных Microsoft Common Object in Context[5]. Он содержит 21 000 изображений, каждое из которых сопровождается как минимум пятью описаниями. Словарь содержит 6000 слов.

Для оценки качества построения тематической модели использовались следующие оценки качества:

1) перплексия, определяемая следующей формулой [34]:

$$P = \exp\left(-\frac{1}{n}\sum_{\mathbf{i}\in I}\sum_{\mathbf{w}\in i} n_{\mathbf{iw}} \ln p(\mathbf{w}|\mathbf{i})\right),$$

   где n — длина коллекции в векторных представлениях слов. Перплексия зависит от мощности словаря и распределения частот слов в коллекции;
2) разреженность матриц Φ и Θ, определяемая как доля нулевых элементов в соответствующих матрицах;
3) контрастность, определяемая следующей формулой:

$$\text{Contrast}_t = \frac{1}{|W_t|}\sum_{\mathbf{w}\in W_t} p(t|\mathbf{W});$$

4) чистота ядра темы, определяемая следующей формулой:

$$\text{Purity}_t = \sum_{\mathbf{w}\in W_t} p(\mathbf{w}|t).$$

[5] `http://mscoco.org/`.

Для оценки качества использования модели для задачи аннотации изображений использовались следующие оценки качества: точность (precision), полнота (recall), а также F_1-мера [35].

Для оценки качества использования модели для задачи иллюстрирования текста использовалась точность (accuracy).

8 Результаты экспериментов

8.1 Качество построения модели

Были проведены эксперименты с использованием разных комбинаций регуляризаторов, а также с разным числом тем и итераций построения модели. Исследования проводились для числа тем, равного 50. Это число было выбрано на основе нескольких экспериментов. Следует отметить, что, как было показано в [36], в реальных задачах не существует оптимального числа тем. Набор регуляризаторов и их параметры подбирались также экспериментально. Наилучшими характеристиками обладает модель с использованием комбинации регуляризаторов разреженности для матриц Φ и Θ и декорреляции тем Φ с коэффициентами $-0{,}013$, $-0{,}25$ и $5{,}2 \cdot 10^5$ соответственно.

Для сравнения была выбрана модель без использования регуляризаторов, имитирующая работу PLSA.

Результаты экспериментов приведены в табл. 1 и на рис. 3. Для большей наглядности на рис. 3, *а* отображается логарифм перплексии.

Как видно из табл. 1, использование регуляризаторов дает прирост по всем видам метрик.

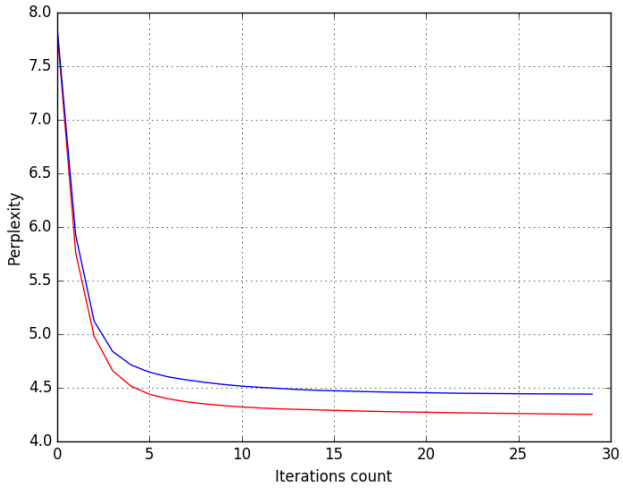
Таблица 1 Сравнение моделей ARTM и PLSA: метрики качества P — перплексия на выборке в 3000 изображений; S_Φ и S_Θ — разреженности матриц Φ и Θ; K_p и K_c — средняя чистота и контрастность ядер тем

Модель	P	S_Φ, %	S_Θ, %	K_p	K_c
ARTM	**70,312**	**96,5**	**88,6**	**0,889**	**0,831**
PLSA	84,597	82,1	84,6	0,461	0,656

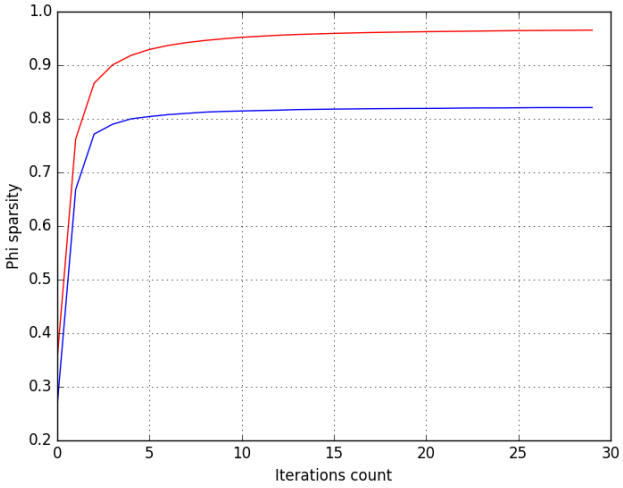
8.2 Аннотация изображений

Для тестирования применялась кросс-валидация на 20% набора данных для тестирования.

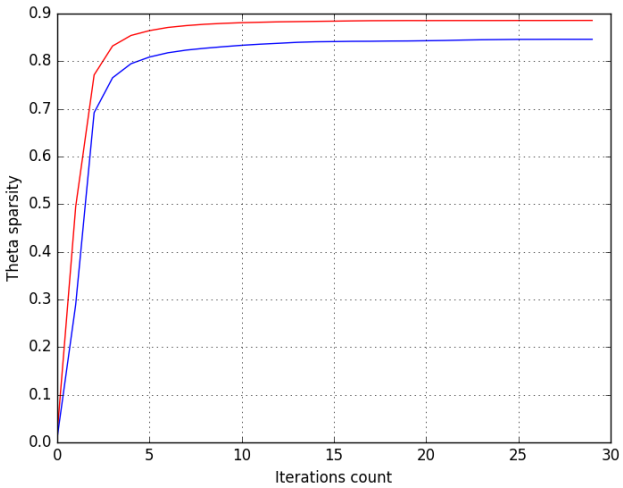
Процесс классификации выглядит следующим образом. На вход классификатору подается изображение без аннотации, на выходе классификатор выдает 10 лучших слов в качестве описания. Для валидации использовалась настоящая аннотация, разбитая на слова.



(*а*)



(*б*)



(*в*)

Рис. 3 График $\log P$ (*а*) и разреженностей матриц Φ (*б*) и Θ (*в*) для моделей PLSA (синие кривые) и ARTM (красные кривые)

Таблица 2 Результаты тестирования моделей в задаче аннотации изображений

Модель	Полнота	Точность	F_1-мера
CORRLDA	34,83	37,85	36,27
MIXLDA	35,20	37,98	36,54
sLDA	35,63	38,46	36,99
PLSA	35,94	38,02	36,92
ARTM	**40,43**	**43,37**	**41,85**

Сравнение проводилось с моделями CorrLDA[6], MixLDA[7] и sLDA[8], описанными в разд. 2. В качестве модели для тестирования использовались обе модели, описанные в предыдущем разделе. Результаты тестирования приведены в табл. 2.

8.3 Иллюстрирование текста

Для иллюстрирования текста на вход подавалось описание и пул изображений кандидатов. Модель находила изображение из пула, которое лучше всего подходит к описанию. Все модели оценивались с помощью лучшего значения точности, которая означает долю успешно подобранных пар в тестовом наборе. Результаты представлены в табл. 3.

Полученные в ходе сравнения результаты показывают, что предложенная модель превосходит существующие аналоги на используемом наборе данных.

Таблица 3 Результаты тестирования моделей в задаче иллюстрирования текста

Модель	Точность
MixLDA	43,5
PLSA	55,8
ARTM	**60,4**

9 Заключение

В данной работе рассмотрена проблема построения мультимодальных тематических моделей текстов и изображений.

В работе были достигнуты следующие результаты.

1. Предложен новый метод построения тематической модели текстов и изображений, учитывающий также контекст слов с помощью их векторного представления.
2. Продемонстрировано применение полученной модели к задачам аннотирования изображений и иллюстрирования текста.

[6]Исходный код для построения модели CorrLDA был взят с сайта `http://home.in.tum.de/~xiaoh/`.

[7]MixLDA моделировался с помощью библиотеки BigARTM.

[8]Исходный код для построения модели sLDA был взят с сайта `http://www.cs.cmu.edu/~chongw/slda/`.

3. Проведены вычислительные эксперименты, показавшие превосходство предложенной модели над ранее известными на конкретном наборе данных.

Стоит отметить, что полученную модель можно использовать в различных задачах, таких как кластеризация текстов и изображений по темам, генерация изображений по описанию и др. Также модель можно легко расширить путем введения дополнительных модальностей для учета различных признаков.

В дальнейшем авторы планируют исследовать влияние методов, использованных для векторного представления, на качество получаемого результата.

Литература

[1] *Yi X., Allan J.* A comparative study of utilizing topic models for information retrieval // European Conference on Information Retrieval, 2009. P. 29–41.

[2] *Rubin T. N., Chambers A., Smyth P., Steyvers M.* Statistical topic models for multi-label document classification // Mach. Learn., 2012. Vol. 88. No. 1-2. P. 157–208.

[3] *Yang M., Hsu W. H.* Hdpauthor: A new hybrid author-topic model using latent Dirichlet allocation and hierarchical Dirichlet processes // 25th Conference (International) on Companion on World Wide Web Proceedings, 2016. P. 619–624.

[4] *Fu Y., Hospedales T. M., Xiang T., Gong S.* Learning multimodal latent attributes // IEEE Trans. Pattern Anal. Machine Intelligence, 2014. Vol. 36. No. 2. P. 303–316.

[5] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: A survey // Frontiers Computer Science China, 2010. Vol. 4. No. 2. P. 280–301.

[6] *Harris Z. S.* Distributional structure // Word, 1954. Vol. 10. No. 2-3. P. 146–162.

[7] *Papadimitriou C. H., Tamaki H., Raghavan P., Vempala S.* Latent semantic indexing: A probabilistic analysis // 17th ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems Proceedings, 1998. P. 159–168.

[8] *Deerwester S. C., Dumais S. T., Furnas G. W., et al.* Computer information retrieval using latent semantic structure, 1989. U.S. Patent 4,839,853.

[9] *Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P.* Numerical recipes in C. The art of scientific computing. — 2nd ed. — Cambridge University Press, 1996. Vol. 2. 994 p.

[10] *Hofmann T.* Probabilistic latent semantic indexing // 22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings, 1999. P. 50–57.

[11] *McLachlan G., Krishnan T.* The EM algorithm and extensions. — 2nd ed. — Hoboken, NJ, USA: Whiley-Interscience, 2007. Vol. 382. 400 p.

[12] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // J. Mach. Learn. Res., 2003. Vol. 3. P. 993–1022.

[13] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Докл. РАН, 2014. Т. 456. №3. С. 268–271.

[14] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-Bayesian additive regularization for multimodal topic modeling of large collections // Workshop on Topic Models: Post-Processing and Applications Proceedings. — New York, NY, USA: ACM, 2015. P. 29–37.

[15] *Frei O., Apishev M.* Parallel non-blocking deterministic algorithm for online topic modeling // Analysis of Images, Social Networks and Texts, 2016.

[16] *Blei D. M., Jordan M. I.* Modeling annotated data // 26th Annual ACM SIGIR Conference (International) on Research and Development in Informaion Retrieval, 2003. P. 127–134.

[17] *Meghini C., Sebastiani F., Straccia U.* A model of multimedia information retrieval // J. ACM, 2001. Vol. 48. No. 5. P. 909–970.

[18] *Chong W., Blei D., Li F.-F.* Simultaneous image classification and annotation // IEEE Conference on Computer Vision and Pattern Recognition, 2009. P. 1903–1910.

[19] *Feng Y., Lapata M.* Topic models for image annotation and text illustration // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010. P. 831–839.

[20] Object recognition from local scale-invariant features // 7th IEEE Conference (International) on Computer Vision Proceedings, 1999. Vol. 2. P. 1150–1157.

[21] *Mcauliffe J. D., Blei D. M.* Supervised topic models // Advances in Neural Information Processing Systems, 2008. P. 121–128.

[22] *Shapiro L., Rosenfeld A.* Computer vision and image processing. — San Diego, CA, USA: Academic Press, 1992. 662 p.

[23] *Krizhevsky A., Sutskever I., Hinton G. E.* Imagenet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems, 2012. P. 1097–1105.

[24] *Athiwaratkun B., Kang K.* Feature representation in convolutional neural networks. arXiv preprint, 2015. arXiv:1507.02313.

[25] *Bengio Y., Ducharme R., Vincent P., Jauvin C.* A neural probabilistic language model // J. Mach. Learn. Res., 2003. Vol. 3. P. 1137–1155.

[26] *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space. arXiv preprint, 2013. arXiv:1301.3781.

[27] *Goldberg Y., Levy O.* Word2Vec explained: Deriving Mikolov *et al.*'s negative-sampling word-embedding method. arXiv preprint, 2014. arXiv:1402.3722.

[28] *Plisson J., Lavrac N., Mladenic D., et al.* A rule based approach to word lemmatization // 7th Multi-Conference (International) Information Society Proceedings, 2004. Vol. 1. P. 83–86.

[29] *Dolamic L., Savoy J.* Stemming approaches for East European languages // Workshop of the Cross-Language Evaluation Forum for European Languages, 2007. P. 37–44.

[30] *Smirnov I.* Overview of stemming algorithms // Mechanical Translation, 2008. Vol. 52.

[31] *Jongejan B., Dalianis H.* Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike // Joint Conference of the 47th Annual Meeting of the ACL and 4th Joint Conference (International) on Natural Language Processing of the AFNLP Proceedings, 2009. Vol. 1. P. 145–153.

[32] *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models // Machine Learn., Special Issue on Data Analysis and Intelligent Optimization with Applications, 2015. Vol. 101. No. 1. P. 303–323.

[33] *Simonyan K., Zisserman A.* Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014. arXiv:1409.1556.

[34] *Brown P. F., Pietra V. J. D., Mercer R. L., Pietra S. A. D., Lai J. C.* An estimate of an upper bound for the entropy of english // Comput. Linguistics, 1992. Vol. 18. No. 1. P. 31–40.

[35] *Friedman J., Hastie T., Tibshirani R.* The elements of statistical learning. — 2nd ed. — Springer ser. in statistics. — Berlin: Springer, 2009. 745 p.

[36] *Vorontsov K. V., Potapenko A. A., Plavin A. V.* Additive regularization of topic models for topic selection and sparse factorization // 3rd Symposium (International) on Learning and Data Sciences. — London, U.K.: University of London, 2015. P. 193–202.

Поступила в редакцию 04.09.2016

Multimodal topic model for texts and images utilizing their embeddings*

N. D. Smelik and A. A. Filchenkov
smelik@rain.ifmo.ru; afilchenkov@corp.ifmo.ru
ITMO University, 49 Kronverksky Pr., St. Petersburg, Russia

A joint topic model for texts and images allows to extract image topics based on their text annotations and to suggest annotations for new images. A novel multimodal topic model for images and texts has been introduced. The proposed model utilizes vector representation of texts and images. Vector representation for a text is based on Word2Vec embedding. Vector representation for an image is convolutional neural network feature map. Then, vector of image is considered as a pseudodocument containing vectors of words instead of words. The proposed model is learnt on the resulting pseudodocument collection. An algorithm to learn the model as well as an algorithm for image annotating and an algorithm for text illustrating with a learnt model have been proposed. Microsoft Common Object in Context dataset was used for experiments. It contains 21,000 images, each has at least 5 annotations. The results show that usage of ARTM leads to much higher quality than the usage of PLSA. The present model was compared with CORRLDA, MIXLDA, and sLDA in image annotating problem and with MIXLDA in text illustrating problem. In both cases, the proposed model showed better results.

Keywords: *topic model; image annotation; text illustration; convolutional neural networks; word embedding*

DOI: 10.21469/22233792.2.4.05

References

[1] Yi, X., and J. Allan. 2009. A comparative study of utilizing topic models for information retrieval. *European Conference on Information Retrieval*. 29–41.

[2] Rubin, T. N., A. Chambers, P. Smyth, and M. Steyvers. 2012. Statistical topic models for multi-label document classification. *Mach. Learn.* 88(1-2):157–208.

[3] Yang, M., W. H. Hsu. 2016. Hdpauthor: A new hybrid author-topic model using latent Dirichlet allocation and hierarchical Dirichlet processes. *25th Conference (International) on Companion on World Wide Web Proceedings*. 619–624.

[4] Fu, Y., T. M. Hospedales, T. Xiang, and S. Gong. 2014. Learning multimodal latent attributes. *IEEE Trans. Pattern Anal. Machine Intelligence* 36(2):303–316.

[5] Daud, A., J. Li, L. Zhou, and F. Muhammad. 2010. Knowledge discovery through directed probabilistic topic models: A survey. *Frontiers Computer Science China* 4(2):280–301.

[6] Harris, Z. S. 1954. Distributional structure. *Word* 10(2-3):146–162.

[7] Papadimitriou, C. H., H. Tamaki, P. Raghavan, and S. Vempala. 1998. Latent semantic indexing: A probabilistic analysis. *17th ACM SIGACT–SIGMOD–SIGART Simposium on Principles of Database Systems Proceedings*. 159–168.

[8] Deerwester, S. C., S. T. Dumais, G. W. Furnas, *et al.* 1989. Computer information retrieval using latent semantic structure. U.S. Patent 4,839,853.

[9] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1996. *Numerical recipes in C. The art of scientific computing*. 2nd ed. Cambridge University Press. Vol. 2. 994 p.

*The research was supported by the Russian Government (grant 074-U01) and the Russian Foundation for Basic Research (project No. 16-37-60115).

[10] Hofmann, T. 1999. Probabilistic latent semantic indexing. *22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings*. 50–57.

[11] McLachlan, G., and T. Krishnan. 2007. The EM algorithm and extensions. 2nd ed. Hoboken, NJ: Whiley-Interscience. Vol. 382. 400 p.

[12] Blei, D. M., A. Y. Ng, and M. I. Jordan. 203. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

[13] Vorontsov, K. V. 2014. Additive regularization for topic models of text collections. *Dokl. Math.* 89(3):301–304.

[14] Vorontsov, K., O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina. 2015. Non-Bayesian additive regularization for multimodal topic modeling of large collections. *Workshop on Topic Models: Post-Processing and Applications Proceedings*. New York, NY: ACM. 29–37.

[15] Frei, O., and M. Apishev. 2016. Parallel non-blocking deterministic algorithm for online topic modeling. *Analysis of Images, Social Networks and Texts*.

[16] Blei, D. M., and M. I. Jordan. 2003. Modeling annotated data. *26th Annual ACM SIGIR Conference (International) on Research and Development in Informaion Retrieval Proceedings*. 127–134.

[17] Meghini, C., F. Sebastiani, and U. Straccia. 2001. A model of multimedia information retrieval. *J. ACM* 48(5):909–970.

[18] Chong, W., D. Blei, and F.-F. Li. 2009. Simultaneous image classification and annotation. *IEEE Conference on Computer Vision and Pattern Recognition*. 1903–1910.

[19] Feng, Y., and M. Lapata. 2010. Topic models for image annotation and text illustration. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 831–839.

[20] Lowe, D. G. 1999. Object recognition from local scale-invariant features. *7th IEEE Conference (International) on Computer Mission*. 2:1150–1157.

[21] Mcauliffe, J. D., and D. M. Blei. 2008. Supervised topic models. *Advances in Neural Information Processing Systems*. 121–128.

[22] Shapiro, L., and A. Rosenfeld. 1992. *Computer vision and image processing*. San Diego, CA: Academic Press. 662 p.

[23] Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 1097–1105.

[24] Athiwaratkun, B., and K. Kang. 2015. Feature representation in convolutional neural networks. arXiv preprint. arXiv:1507.02313

[25] Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–1155.

[26] Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint. arXiv:1301.3781.

[27] Goldberg, Y., and O. Levy. 2014. Word2Vec explained: Deriving Mikolov *et al.*'s negative-sampling word-embedding method. arXiv preprint. arXiv:1402.3722.

[28] Plisson, J., N. Lavrac, D. Mladenic, *et al.* 2004. A rule based approach to word lemmatization. *7th Multi-Conference (International) Information Society Proceedings*. 1:83–86.

[29] Dolamic, L., and J. Savoy. 2007. Stemming approaches for East European languages. *Workshop of the Cross-Language Evaluation Forum for European Languages*. 37–44.

[30] Smirnov, I. 2008. Overview of stemming algorithms. *Mechanical Translation* 52.

[31] Jongejan, B., and H. Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. *Joint Conference of the 47th Annual Meeting of the ACL and 4th Joint Conference (International) on Natural Language Processing of the AFNLP Proceedings*. 1:145–153.

[32] Vorontsov, K. V., and A. A. Potapenko. 2015. Additive regularization of topic models. *Machine Learn.* Special Issue on Data Analysis and Intelligent Optimization with Applications 101(1):303–323.

[33] Simonyan, K., and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556.

[34] Brown, P. F., V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, and J. C. Lai. 1992. An estimate of an upper bound for the entropy of english. *Comput. Linguistics* 18(1):31–40.

[35] Friedman, J., T. Hastie, and R. Tibshirani. 2009. *The elements of statistical learning*. 2nd ed. Springer ser. in statistics. Berlin: Springer. 745 p.

[36] Vorontsov, K. V., A. A. Potapenko, and A. V. Plavin. 2015. Additive regularization of topic models for topic selection and sparse factorization. *3rd Symposium (International) on Learning and Data Sciences*. London, U.K.: University of London. 193–202.

Received September 4, 2016

Метод определения положения век на изображении при распознавании человека по радужной оболочке глаза с мобильного устройства

***Г. А. Одиноких*[1,2], *В. С. Гнатюк*[1,2], *М. В. Коробкин*[1,2], *В. А. Еремеев*[1,2]**
g.odinokikh@gmail.com; vitgracer@gmail.com; mikhail.korobkin@hotmail.com
[1]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2
[2]Национальный исследовательский университет МИЭТ
Россия, г. Москва, г. Зеленоград, пл. Шокина, 1

При распознавании человека по радужке информация о положении век на изображении используется для удаления шума от век и ресниц, перекрывающих полезную область радужки, оценки качества изображения и многих других целей. Определение положения век, как правило, производится после вычислительно сложной операции нахождения границ радужки и склеры. В случае использования для распознавания мобильного устройства такой подход не всегда оправдан ввиду, в частности, ограниченной производительности устройства, сложностей взаимодействия пользователя с устройством и сильно изменяющихся внешних условий окружающей среды. В данном случае информация о положении век может быть извлечена сразу после этапа детектирования зрачка и использована для определения пригодности изображения для последующих более сложных этапов алгоритма распознавания. Предложен метод определения положения век на изображении с целью оценки качества изображения и последующего определения границы радужки и века. Производительность метода была оценена в сравнении с несколькими существующими решениями с использованием четырех различных открытых баз данных изображений радужек.

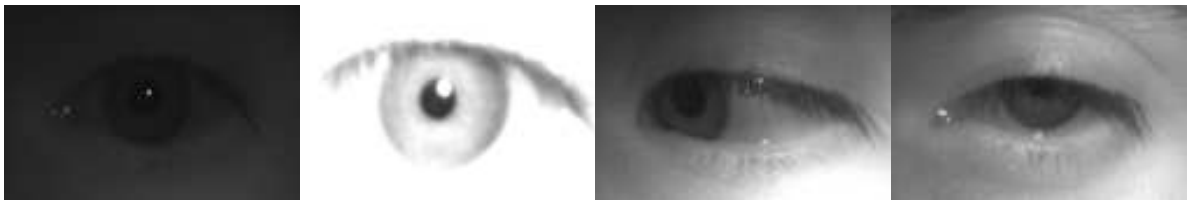
Ключевые слова: *оценка качества изображения; мобильная биометрия; распознавание по радужке*

DOI: 10.21469/22233792.2.4.06

1 Введение

Биометрические технологии идентификации личности хорошо зарекомендовали себя при использовании в различных сферах человеческой деятельности, в частности в системах контроля и управления доступом. Такая тенденция связана с постоянным повышением требований к уровню безопасности и удобству при проведении аутентификации пользователя. Одним из ярких примеров, на сегодняшний день, является применение биометрических технологий в мобильных устройствах, в частности при проведении платежных операций и ограничении доступа к персональным данным пользователя устройства. Здесь методы биометрической аутентификации рассматриваются как достойный кандидат, приходящий на смену традиционным, таким как пароли, смарт-карты, ПИН-коды и т. д. Биометрические технологии, использующие изображение радужной оболочки глаза в качестве биометрического признака, обладают рядом преимуществ по сравнению с остальными [1–3], что делает их особенно привлекательными для использования в мобильных устройствах.

Использование биометрической системы в мобильном устройстве подразумевает ее способность обрабатывать биометрические данные в условиях постоянно изменяющейся окружающей среды и удобство взаимодействия с пользователем. Изменение условий



(*а*) (*б*) (*в*) (*г*)

Рис. 1 Примеры изображений радужной оболочки глаза, полученные с мобильного устройства: (*а*) низкая освещенность; (*б*) переэкспонирование; (*в*) отвод взгляда; (*г*) затенение веком и ресницами

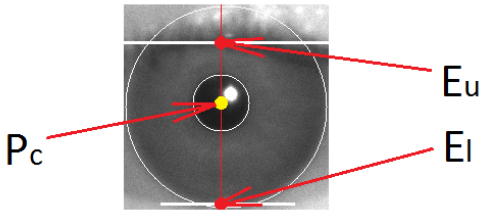


Рис. 2 Определение положения века: E_u и E_l — точки, соответствующие положениям верхнего и нижнего век; P_c — центр зрачка

среды влечет за собой значительное ухудшение качества входных биометрических данных (изображения, в случае с радужкой) и, как следствие, оказывают сильное влияние на производительность системы распознавания [4]. На ухудшение качества изображения также оказывают сильное влияние и сложности взаимодействия с пользователем, например частое моргание, дрожание рук, отвод взгляда и т. д. Некоторые примеры изображений радужки, полученные с мобильного устройства, представлены на рис. 1. Помимо вышеперечисленного система должна обеспечивать распознавание в режиме реального времени на устройстве с ограниченными вычислительными ресурсами и объемом памяти.

В данной работе предлагается метод определения положения век на изображении при распознавании человека по радужной оболочке глаза. Положение каждого века определяется как минимальное расстояние от границы века до центра зрачка (рис. 2).

Значения E_u и E_l могут быть использованы для оценки качества входящего биометрического образца, в частности степени открытости глаза. На первом этапе эта информация может быть использования для построения классификатора либо решающего правила об отсеивании образца, непригодного для распознавания, а также для предоставления обратной связи с пользователем (отображения подсказки). В том случае если образец был классифицирован как пригодный для дальнейшей обработки, значения E_u и E_l могут быть использованы на дальнейших этапах распознавания, таких как определение границ радужка–склера, радужка–веко, а также на этапе выделения ресниц.

Большинство существующих методов рассматривают процедуру единовременного определения полной границы века либо границы радужка–веко, и ни один из них не рассматривает обработку в два этапа, как предлагается в данной работе. Поэтому для сравнения с существующими методами, точки E_u и E_l для них были восстановлены сразу после нахождения полной границы век.

Предлагаемый в работе метод позволяет с высокой точностью определять положение век на изображении сразу после этапа нахождения зрачка. Метод обладает высокой

устойчивостью и скоростью обработки и, таким образом, может быть применим в системах распознавания по радужке на мобильном устройстве.

2 Обзор существующих методов

В алгоритмах распознавания человека по радужке этап определения положения век обычно следует за этапом определения границы радужка–склера [1, 5–8] либо после этапа нормализации радужки (рис. 3) [9]. Оба этих этапа требуют дополнительных вычислений. В данной работе рассматриваются только те из методов, которые предусматривают детектирование век после определения положения зрачка или радужки.

Структура алгоритма детектирования век на изображении может быть поделена на две основные части: предобработка изображения и локализация век. Зачастую исследователи в своих работах предлагают лишь одну из частей, а в качестве второй берут существующий подход. В данной работе произведена оценка существующих комбинаций предобработка–локализация, а также нескольких возможных комбинаций, ранее не описанных в литературе. Существующие методы и комбинации, продемонстрировавшие наилучшие результаты по точности детектирования век, рассмотрены в сравнении с предложенным методом.

Для нахождения границы века Дж. Дугман предложил сглаживание изображения с использованием фильтра Гаусса и применение интегродифференциального оператора (IDO) [1] для поиска параболических кривых. Уайлдс в своей работе [10] предложил выделение границ на изображении в качестве первого шага и преобразование Хо для последующей локализации века. Масек предложил разделять области радужки и века горизонтальными прямыми [11]. Положение прямых вдоль вертикальной оси определялось в соответствии со значением максимального отклика после свертки исходного изображения с фильтром Собела в регионе поиска. Такую же свертку использовали в своей работе Канг и Парк [12], дополнив ее IDO для уточнения. Сианде и др. в своей работе [6] предложили использование одномерного фильтра пиковой формы для удаления шума от ресниц на изображении и IDO для параболических кривых для локализации. Адам и др. в работах [5, 13] применили анизотропную диффузию для подавления шума, фильтр Собела для выделения границ, а затем преобразование Хо для параболических кривых. Янг и др. предложили использовать асимметричный оператор Кэнни для выделения границ века и аппроксимацию параболической кривой методом наименьших квадратов для локализации [14]. Ким и др. использовали информацию о положении век на изображении для определения направления взгляда [15], применив для этого выравнивание гистограммы в качестве предобработки и поиск локального минимума для определения положения века. Хе и др. предложили использование одномерного нелинейного ранг-фильтра для удаления шума от ресниц и статистическую модель кривизны века для локализации [16].

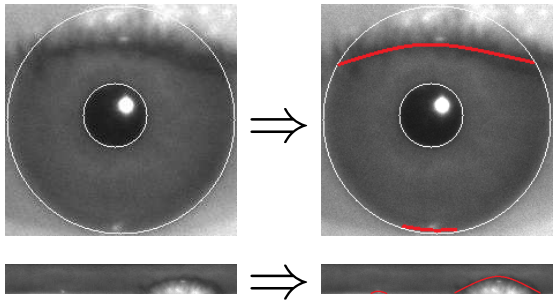


Рис. 3 Этапы определения положения век

Почти все вышеупомянутые методы были предложены и протестированы их авторами исключительно на базах данных изображений, полученных не с помощью мобильного устройства, и, таким образом, не предусматривают возможность устойчивой работы в постоянно изменяющихся условиях окружения, упомянутых ранее.

3 Определение положения век на изображении

В данной работе предложен метод определения положения век на изображении для применения его, в частности, в задачах биометрического распознавания человека по радужной оболочке глаза. Информация о положении век может быть использована для оценки степени открытости глаза, а также на последующих этапах локализации границ века, радужки и при поиске ресниц. Оценка степени открытости глаза производится сразу после этапа детектирования зрачка на изображении. Предложенный метод основан на применении многонаправленного двумерного (2D) фильтра Габора и одномерного (1D) выборочного извлечения границ в качестве этапа предобработки, а также метода скользящего окна для детектирования положения века (локализации) (рис. 4)

Рис. 4 Алгоритм детектирования век на изображении

Свертка с ядром 2D фильтра Габора в качестве метода предобработки и его параметры были выбраны из следующих соображений: граница радужка–веко представляет собой кривую, разделяющую две области различной интенсивности; область границы также часто характеризуется наличием тени от века, которая становится все более различимой с повышением уровня освещенности в помещении; параметры фильтра были подобраны таким образом, чтобы подчеркнуть границу радужка–веко, используя как информацию о самой границе, так и информацию о тени от века; ориентация θ и количество ядер свертки были выбраны с учетом возможной ориентации века, формы века, а также шума, вызванного ресницами и тенями от ресниц. Метод одномерного выборочного извлечения границ использовался для подавления шума, вызванного иными различными текстурами области века: складки кожи, ориентированные горизонтально ресницы и т. п. Совокупность использованных методов делает предложенное решение устойчивым к изменениям окружающей среды и позволяет с высокой точностью определять положение век даже в сложных случаях, связанных с особенностями поведения пользователя.

В соответствии с диаграммой (см. рис. 4) предлагаемый алгоритм может быть представлен в более наглядной форме (рис. 5).

3.1 Этап предобработки изображения

В качестве входных данных для алгоритма используются данные о зрачке: координаты его центра X_p и Y_p и значение радиуса R_p. Параметры области интереса определены

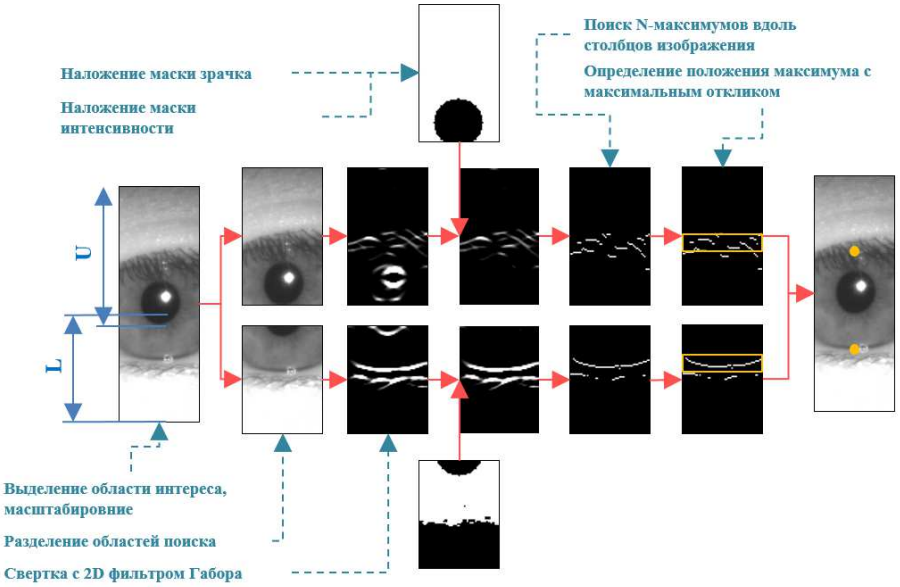


Рис. 5 Алгоритм детектирования век на изображении

следующим образом: $\text{width} = n_W R_p$; $\text{height} = n_H R_p$, где n_W и n_H — константы, определяющие ширину и высоту изображения региона интереса соответственно. В данной работе были выбраны значения $n_W = 4$ и $n_H = 12$; таким образом, $n_H/n_W = 3$. Данные значения были подобраны экспериментально и позволяют учитывать отклонение положения века от закрытого до полностью открытого состояния, а также ограничивать область поиска по горизонтали. Несмотря на то что радиус зрачка может значительно изменяться в зависимости от множества факторов, данный подход продемонстрировал высокую точность определения положения века (см. табл. 1 и 2 в разд. 4) для всех баз данных, использованных при тестировании.

После выделения области интереса над изображением осуществляется операция масштабирования с коэффициентом 0,5. Масштабирование осуществляется исключительно в целях ускорения обработки данных. Экспериментально показано, что данное значение коэффициента является достаточным для обеспечения высокой скорости обработки данных без деградации точности детектирования.

Следующим этапом предобработки является разделение изображения на две области: верхнего и нижнего век соответственно. Разделение производится с использованием следующих правил: $U \in (0, Y_p + R_p)$ и $L \in (Y_p + R_p/2, Y_{\max})$, как представлено на рис. 5, где 0 соответствует первой, а $Y_{\max}$ равен числу строк изображения.

Далее для каждого из изображений производится операция свертки с ядром двумерной функции Габора вида:

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(-i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right).$$

Здесь

$$x' = x\cos\theta + y\sin\theta\,; \quad y' = -x\sin\theta + y\cos\theta\,;$$

$\lambda, \theta, \psi, \sigma$ и γ — длина волны, ориентация, значение фазы, величина стандартного отклонения Гауссового ядра и коэффициент сжатия Гауссового ядра соответственно.

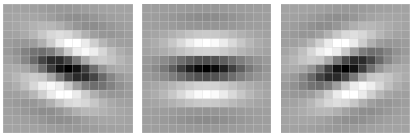


Рис. 6 Габоровские ядра различных ориентаций (θ_n)

Свертка с фильтром Габора производится для обоих изображений (верхнего и нижнего век) для N_o различных ориентаций (θ_n). В данной работе $N_o = 3$ и ориентации для изображений верхнего и нижнего век (град) соответственно равны: $\theta_{1\ldots N}^{\text{Upper}} = 250{,}0$, 270,0 и 290,0 и $\theta_{1\ldots N}^{\text{Lower}} = 70{,}0$, 90,0 и 110,0.

Приведенные выше ориентации были также получены экспериментально, так как с их использованием достигнута наивысшая точность детектирования. Над каждым из изображений век $I(x, y)$ осуществляется операция свертки с заданным ядром:

$$I'(x, y) = I(x, y) \otimes g'(i, j)\,.$$

Ядро $g'(i, j)$ является скомбинированным из N ядер фильтра Габора различных ориентаций θ_n соответственно:

$$g'(i, j) = \frac{1}{N} \sum_{n=1}^{N} g(i, j, \theta_n)\,.$$

Примеры ядер различных ориентаций θ_n, использованных для свертки, представлены на рис. 6. Изображения $I'(x, y)$, полученные после свертки, представлены на схеме (см. рис. 5).

В соответствии со схемой (см. рис. 5) на следующем шаге производится наложение маски $M(x, y)$, скомбинированной из двух: маски зрачка $M_P(x, y)$ и маски интенсивности $M_I(x, y)$. Маска может быть получена следующим образом:

$$M(x, y) = M_p(x, y) \wedge M_I(x, y)\,.$$

Маска зрачка $M_P(x, y)$ определяет область зрачка на изображении, а маска интенсивности $M_I(x, y)$ — переэкспонированную область изображения:

$$M_I(x, y) = \begin{cases} 255, & \text{если } I_s(x, y) < 240; \\ 0 & \text{иначе,} \end{cases}$$

где $I_s(x, y)$ — результат применения к исходному изображению века $I(x, y)$ фильтра Гаусса. Параметры ядра: size $= 3 \times 3, \sigma = 1{,}0$. Над комбинированной маской $M(x, y)$ также производится операция дилатации на величину, равную половине размера ядра Гауссиана.

Последним этапом предобработки является извлечение карты границ. Извлечение производится посредством поиска N_p максимальных значений градиента вдоль каждого из столбцов изображения $I'(x, y)$ в направлении от нижней до верхней границы для изображения верхнего века и в обратном направлении для изображения нижнего века (см. рис. 5). В данной работе $N_p = 2$, а размер окна для вычисления градиента равен 3.

3.2 Этап локализации век

Для определения положения век использован метод скользящего окна (см. рис. 5). Функция отклика отражает зависимость координаты центра окна от количества точек границ, попадающих внутрь окна, вдоль вертикальной оси.

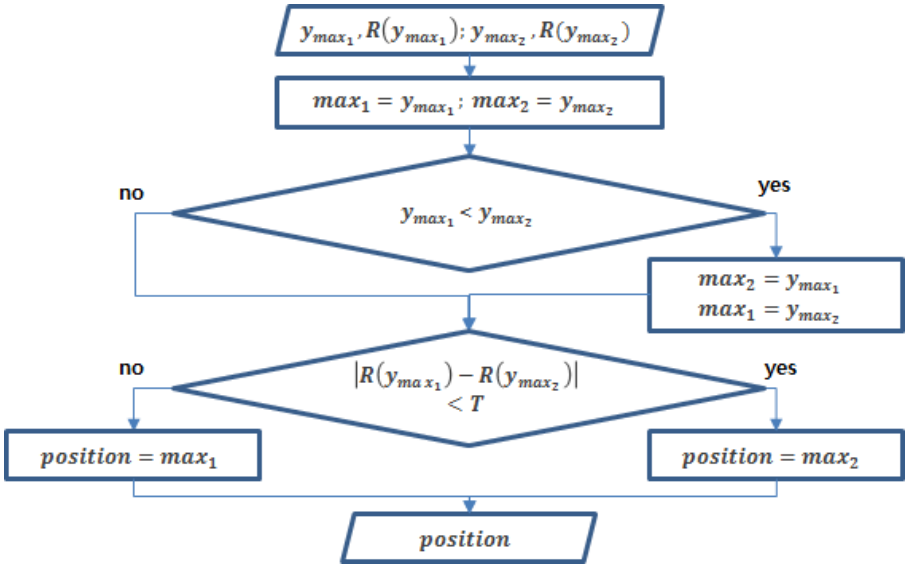


Рис. 7 Решающее правило для точек кандидатов (T — значение порога; в данной работе выбрано $T = R(y_{\max 1})/3$)

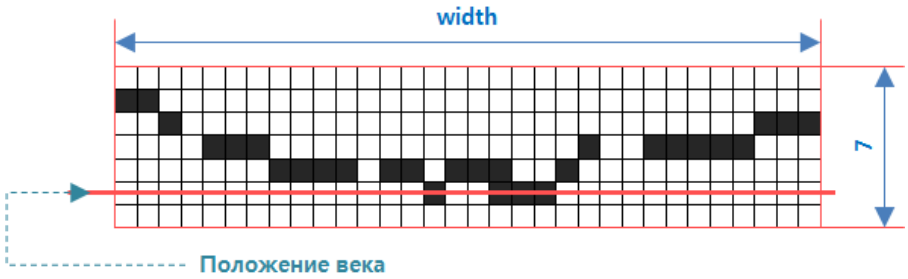


Рис. 8 Определение положения века внутри окна

Далее два положения окна ($y_{\max 1}, y_{\max 2}$), соответствующие максимальным значениям функции отклика $R(y)$, используются в качестве кандидатов. Для выбора между ними применяется решающее правило вида (рис. 7).

Подход с использованием двух максимумов ($N_p = 2$) в совокупности с применением решающего правила (см. рис. 7) для двух точек-кандидатов позволил значительно уменьшить количество ошибок детектирования, вызванных различными шумовыми факторами: ресницами, складками кожи, дужками очков и др.

После того как финальное положение окна выбрано, определение положения века осуществляется внутри него. Как изображено на рис. 8, положение века соответствует положению точки границы, оказавшейся внутри окна, максимально удаленной от центра зрачка.

4 Экспериментальные результаты

4.1 Точность детектирования

В целях демонстрации производительности предлагаемого метода были реализованы более 10 различных существующих методов. Однако, для сравнения использовались лишь те из них, которые показали наилучшие результаты по точности детектирования. Все методы протестированы на четырех разных базах данных: CASIA-IrisV4-Thousand [17],

Таблица 1 Точность определения E_u — верхнего века (%), $\xi^{\text{adm}} = 5\%$

База данных	MIR	CS4	CS3	APX	AVG
Дугман [1]	76	70	83	84	74,4
Уайлдс [10]	80	83	92	74	80,6
Масек [11]	50	70	90	93	72,6
Канг и Парк [12]	86	89	90	88	86,0
Сианде и др. [6]	56	92	95	94	83,2
Адам и др. [5]	80	83	91	78	81,2
Янг и др. [14]	55	83	78	90	72,4
Ким и др. [15]	89	89	99	98	89,0
Хе и др. [16]	80	83	92	74	80,6
2DGF+IDO	93	90	95	91	92,3
Предложенный метод	98	97	97	91	94,8

Таблица 2 Точность определения E_l — нижнего века (%), $\xi^{\text{adm}} = 5\%$

База данных	MIR	CS4	CS3	APX	AVG
Дугман [1]	88	86	95	94	90,8
Уайлдс [10]	87	78	92	92	87,3
Масек [11]	40	65	86	95	71,5
Канг и Парк [12]	96	88	95	94	93,3
Сианде и др. [6]	77	87	87	92	85,8
Адам и др. [5]	87	79	93	95	88,5
Янг и др. [14]	12	28	34	72	36,5
Ким и др. [15]	30	50	22	32	33,5
Хе и др. [16]	87	78	92	92	87,3
2DGF+IDO	97	86	92	96	92,8
Предложенный метод	99	94	96	94	95,8

CASIA-IrisV3-Lamp [18], AOPTIX [19] и MIR2016 (Train) — база данных, полученная при помощи бильного устройства [20]. В целях компактного представления результатов эксперимента, в табл. 1 и 2 для баз данных использовались сокращенные обозначения CS4, CS3, APX и MIR соответственно. Для оценки результатов по точности детектирования более 500 изображений для каждой из баз данных были размечены вручную экспертом. Маркировка производилась для изображений, соответствующим области интереса после масштабирования (см. рис. 5). Для каждого из изображений размечались y-координаты точек E_u и E_l (см. рис. 2).

Точность детектирования век определялась для различных значений допустимой ошибки. Выбор такого метода оценивания оправдан в случае, когда необходимо достижение определенного уровня точности для заданного абсолютного значения допустимой ошибки детектирования. Точность определения положения век при распознавании человека по радужке играет существенную роль, в частности в связи с тем, что некоторые последующие этапы алгоритма распознавания используют информацию о положении век.

Точность детектирования оценивалась для трех различных значений допустимой ошибки: ξ_j^{adm}:

$$\xi_{1\ldots N_e}^{\text{adm}} = \{5\%, 10\%, 15\%\}\,.$$

Относительная ошибка ε_j для каждого ξ_j^{adm} и для каждой из баз данных вычислена как мощность множества изображений, для которых ошибка детектирования (в пикселах) превышает пороговое значение допустимой ошибки ξ_j^{adm}:

$$\varepsilon_j = \frac{1}{N} \left| \left\{ \forall i : d_i > \xi_j^{\text{adm}} * \text{height} \right\} \right|,$$

где

$$d_i = \left| E(x, y_A)_i - E(x, y_M)_i \right|;$$

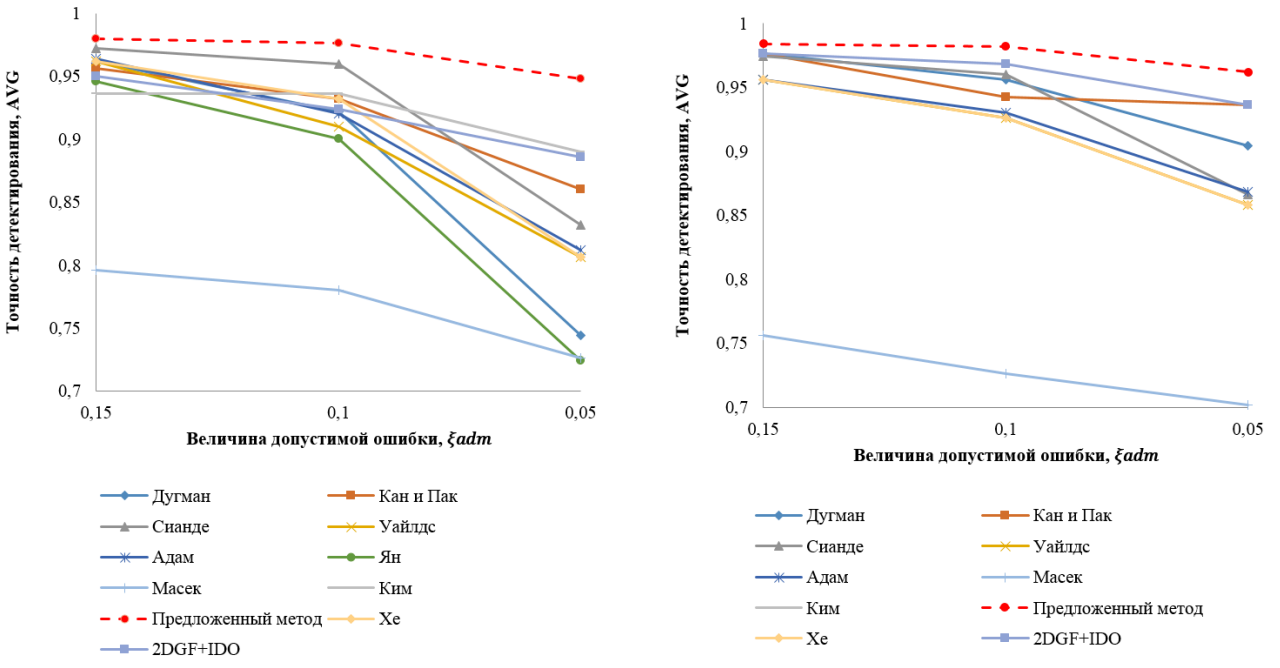
$E(x, y_A)_i$ и $E(x, y_M)_i$ — положения век, олученные по результатам применения алгоритма (A) и размеченные вручную (M) для i-го изображения каждой из баз данных; height — высота изображения, соответствующего региону интереса (до масштабирования) в пикселах; N — количество изображений в базе данных.

В целях демонстрации устойчивости метода для различных условий, а также компактного представления результатов значение точности усреднялось для всех баз данных:

$$\text{AVG}_{\xi_j^{\text{adm}}} = \frac{100\%}{N_D} \sum_{i=1}^{N_D} \left(1{,}0 - \varepsilon_j^i\right), \tag{1}$$

где i — индекс базы данных; N_D — количество использованных баз данных ($N_D = 4$).

Зависимости между значениями AVG и ξ_j^{adm} для различных методов для верхнего и нижнего век получены в соответствии с (1) и представлены на рис. 9. В дополнение к существующим методам оценка была также произведена для некоторых комбинаций предобработка–локализация, ранее не описанных в литературе. В табл. 1 и 2 представлена детальная информация по точности методов, полученная описанным выше способом (1) для различных баз данных, для значения допустимой ошибки $\xi^{\text{adm}} = 5\%$.



(a) $(б)$

Рис. 9 Точность определения E_u (a) и E_l $(б)$

Методы, показавшие наилучшие результаты по точности детектирования, представлены на рис. 9. Также на графиках (см. рис. 9) показаны результаты для метода, представляющего собой комбинацию многонаправленной Габоровской фильтрации в качестве предобработки изображения и интегродифференциальный оператор Дугмана для локализации века (метод обозначен как 2DGF+IDO).

4.2 Скорость обработки

Оценка по скорости обработки предлагаемого метода получена при помощи мобильного устройства Samsung Galaxy Tab Pro 8.4, Snapdragon 800 CPU (2.26 GHz Quad-core). Время выполнения оценивалось как разница между этапом получения данных о зрачке (X_p, Y_p и R_p) и получением значений E_u и E_l. Измерения проводились на одном ядре устройства. Полное время выполнения составило 1,5–4 мс, что говорит о возможности использования метода в режиме реального времени на мобильном устройстве. Измерения не производились для остальных методов, так как это потребовало бы их дополнительной оптимизации и не гарантировало бы достижение результатов, сопоставимых с результатами авторов.

5 Заключение

В работе предложен метод определения положения века в применении к решению задачи распознавания человека по радужной оболочке глаза. В отличие от существующих методов, использующих информацию о границе радужка–склера в качестве входной, а также производящих поиск полной границы между радужкой и веком, предложенный метод может быть использован для быстрой оценки степени открытости глаза сразу после этапа нахождения зрачка. На основании полученной информации может быть принято решение об отсеивании текущего кадра и необходимости обратной связи с пользователем (отображения подсказки) в случае необходимости. Кроме этого, информация о положении век может быть использована на дальнейших этапах алгоритма распознавания, например для адаптивной подстройки параметров алгоритмов сегментации радужки и др. Метод продемонстрировал высокую точность детектирования (см. табл. 1 и 2) по сравнению с существующими аналогами), а также способность работы на мобильном устройстве в режиме реального времени. Устойчивость метода продемонстрирована путем его тестирования на различных базах данных.

Литература

[1] *Daugman J.* How iris recognition works // IEEE Trans. Circ. Syst. Vid., 2004. Vol. 14. No. 1. P. 21–30.

[2] *Chowhan S., Cocsit L., Shinde G.* Iris biometrics recognition application in security management // Image and Signal Processing Conference Proceedings, 2008. Vol. 1. P. 661–665.

[3] *Bhattacharya V., Mali K.* Iris as a biometric feature: Application // Recogn. Advantages Shortcomings Int. J. Adv. Res., 2013. Vol. 3. No. 6. P. 1410–1415.

[4] *Dorairaj V., Schmidt N., Fahmy G.* Performance evaluation of non-ideal iris based recognition system implementing global ica encoding // Conference (International) on Image Processing Proceedings, 2004. Vol. 3. P. 11–14.

[5] *Adam M., Rossant F., Amiel F., Mikovikova B., Ea T.* Reliable eyelid localization for iris identification // Advanced Concepts for Intelligent Vision Systems Conference Proceedings, 2008. P. 1062–1070.

[6] *Xiangde Z, Qi W., Hegui Z., et al.* Noise detection of iris image based on texture analysis // Chinese Control and Decision Conference Proceedings, 2009. P. 2366–2370.

[7] *Gankin K., Cheusev A., Matveev I.* Iris image segmentation based on approximate methods with subsequent refinements // J. Comput. Syst. Sci. Int., 2014. Vol. 53. No. 2. P. 224–238.

[8] *Solomatin I., Matveev I.* Detecting visible areas of iris by qualifier of local textural features // J. Machine Learning Data Anal., 2016. Vol. 1. No. 14. P. 1919–1929.

[9] *Min T., Park R.* Comparison of eyelid and eyelash detection algorithms for performance improvement of iris recognition // Conference (International) on Image Processings Proceedings, 2008. P. 257–260.

[10] *Wildes R.* Iris recognition an emerging biometric technology // Proc. IEEE, 1997. Vol. 85. No. 9. P. 1348–1363.

[11] *Masek L.* Recognition of human iris patterns for biometric identification // Measurement, 2003. Vol. 32. No. 8. P. 1502–1516.

[12] *Kang B., Park K.* A robust eyelash detection based on iris focus assessment // Pattern Recogn. Lett., 2007. Vol. 28. No. 13. P. 1630–1639.

[13] *Adam M., Rossant F., Amiel F., Mikovikova B., Ea T.* Eyelid localization for iris identification // Radioengineering, 2008. Vol. 17. No. 4. P. 82–85.

[14] *Yang L., Wu T., Dong Y., Fei L.* Eyelid localization using asymmetric canny operator // Conference (International) on Computer Design and Applications Proceedings, 2010. P. 533–535.

[15] *Kim H., Cha J., Lee W.* Eye detection for gaze tracker with near infrared illuminator // Conference (International) on Computational Science and Engineering Proceedings, 2014. P. 458–464.

[16] *He Z., Tan T., Sun Z., Qiu X.* Robust eyelid eyelash and shadow localization for iris recognition // Conference (International) on Image Processing Proceedings, 2008. P. 265–268.

[17] Casia. Casia iris image database v4.0. http://biometrics.idealtest.org/dbDetailForUser.do?id=4.

[18] Casia. Casia iris image database v3. http://biometrics.idealtest.org/dbDetailForUser.do?id=3.

[19] AOptix. Aoptix iris database. http://www.aoptix.com.

[20] CASIA, NLPR. The BTAS competition on mobile iris recognition, 2016. http://biometrics.idealtest.org/2016/MIR2016.jsp.

Поступила в редакцию 01.09.2016

Method of eyelid detection on image for mobile iris recognition

***G. A. Odinokikh*[1,2], *V. S. Gnatyuk*[1,2], *M. V. Korobkin*[1,2], *and V. A. Eremeev*[1,2]**

g.odinokikh@gmail.com; vitgracer@gmail.com; mikhail.korobkin@hotmail.com

[1]Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

[2]National Research University of Electronic Technology, 1 Shokin Sq., Zelenograd, Moscow, Russia

Eyelid detection is a very important part of iris recognition procedure. It is required for further eyelid noise removal and iris image quality estimation. Use of a mobile device imposes additional restrictions on iris recognition performance. In addition to the computational load and memory limitations, the recognition should have real-time performance and consider all the user interaction and changing environment conditions difficulties. A method for fast eyelid position detection for iris image quality estimation and further precise eyelid border localization is proposed. The performance of the proposed method is compared with eight the most reliable eyelid detection methods on four open datasets.

Keywords: *eyelid detection; iris recognition; mobile biometrics*

DOI: 10.21469/22233792.2.4.06

References

[1] Daugman, J. 2004. How iris recognition works. *IEEE Trans. Circ. Syst. Vid.* 14(1):21–30.

[2] Chowhan, S., L. Cocsit, and G. Shinde. 2008. Iris biometrics recognition application in security management. *Image and Signal Processing Conference Proceedings*. 1:661–665.

[3] Bhattacharya, V., and K. Mali. 2013. Iris as a biometric feature: Application. *Recogn. Advantages Shortcomings Int. J. Adv. Res.* 3(6):1410–1415.

[4] Dorairaj, V., N. Schmidt, and G. Fahmy. 2004. Performance evaluation of non-ideal iris based recognition system implementing global ica encoding. *Conference (International) on Image Processing Proceedings*. 3:11–14.

[5] Adam, M., F. Rossant, F. Amiel, B. Mikovikova, and T. Ea. 2008. Reliable eyelid localization for iris identification. *Advanced Concepts for Intelligent Vision Systems Conference Proceedings*. 1062–1070.

[6] Xiangde, Z., W. Qi, Z. Hegui, *et al.* 2009. Noise detection of iris image based on texture analysis. *Chinese Control and Decision Conference Proceedings*. 2366–2370.

[7] Gankin, K., A. Cheusev, and I. Matveev. 2014. Iris image segmentation based on approximate methods with subsequent refinements. *J. Comput. Syst. Sci. Int.* 53(2):224–238.

[8] Solomatin, I., and I. Matveev. 2016. Detecting visible areas of iris by qualifier of local textural features. *J. Machine Learning Data Anal.* 1(14):1919–1929.

[9] Min, T., and R. Park. 2008. Comparison of eyelid and eyelash detection algorithms for performance improvement of iris recognition. *Conference (International) on Image Processings Proceedings*. 257–260.

[10] Wildes, R. 1997. Iris recognition an emerging biometric technology. *Proc. IEEE* 85(9):1348–1363.

[11] Masek, L. 2003. Recognition of human iris patterns for biometric identification. *Measurement* 32(8):1502–1516.

[12] Kang, B., and K. Park. 2007. A robust eyelash detection based on iris focus assessment. *Pattern Recogn. Lett.* 28(13):1630–1639.

[13] Adam, M., F. Rossant, F. Amiel, B. Mikovikova, and T. Ea. 2008. Eyelid localization for iris identification. *Radioengineering* 17(4):82–85.

[14] Yang, L., T. Wu, Y. Dong, and L. Fei. 2010. Eyelid localization using asymmetric Canny operator. *Conference (International) on Computer Design and Applications Proceedings*. 533–535.

[15] Kim, H., J. Cha, and W. Lee. 2014. Eye detection for gaze tracker with near infrared illuminator. *Conference (International) on Computational Science and Engineering Proceedings*. 458–464.

[16] He, Z., T. Tan, Z. Sun, and X. Qiu. 2008. Robust eyelid eyelash and shadow localization for iris recognition. *Conference (International) on Image Processing Proceedings*. 265–268.

[17] Casia. Casia iris image database v4.0. Available at: http://biometrics.idealtest.org/dbDetailForUser.do?id=4 (accessed November 4, 2016).

[18] Casia. Casia iris image database v3. Available at: http://biometrics.idealtest.org/dbDetailForUser.do?id=3 (accessed November 4, 2016).

[19] AOptix. Aoptix iris database. Available at: `http://www.aoptix.com` (accessed July 1, 2014).

[20] CASIA, NLPR. 2016. The BTAS competition on mobile iris recognition. Available at: http://biometrics.idealtest.org/2016/MIR2016.jsp (accessed November 4, 2016).

Received September 1, 2016

Выбор решений при распознавании эмоций по речи

В. П. Кальян
vkalyan@mail.ru
ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Описывается выбор решений в системе распознавания эмоционального состояния человека по речи. Анализируется информативность измерительной базы распознавания на основании паралингвистических, артикуляционных и экстралингвистических особенностей речи с учетом индивидуальных эмоционально-смысловых коннотаций в речи испытуемого, описываются алгоритмы распознавания эмоций по речи, осуществляется выбор из множества решений и их верификация в отношении искренности и правдивости говорящего с учетом ситуативного контекста.

Ключевые слова: *распознавание эмоций; эмоциональная речь; древо принятия решений; пространство речевых признаков; паралингвистические особенности речи; артикуляционные модели; спектральная динамика; речевые форманты; частота основного тона; высота голоса*

DOI: 10.21469/22233792.2.4.07

1 Введение

Возможность определения эмоционального состояния говорящего по речи имеет большое практическое значение. Эмоции естественным образом сопутствуют речи, являясь особым каналом общения, по которому непосредственно передается отношение говорящего к текущей ситуации и содержанию сказанного. Это отношение невольно проявляется в характере речи — данное простое соображение, казалось бы, позволяет утверждать, что построение системы, связывающей параметры речи с ее искренностью и правдивостью, является непреложным фактом, а такие бесконтактные детекторы лжи скоро заменят существующие уже много лет в криминалистической практике полиграфы.

И действительно, уже более 40 лет на рынке услуг по детекции лжи под разными названиями появляются коммерческие приборы, позиционирующие сами себя как «анализаторы стресса в голосе». В 1972 г. в США был получен первый патент на прибор под названием PSE (Psychological Stress Evaluator), чуть позже другими разработчиками был выпущен в свет VSE (Voice Stress Evaluator). Утверждалось, что эти и подобные им приборы (Mark-II, ESM, Hagoth и др.) в отличие от полиграфа способны устанавливать неискренность без подключения к телу человека датчиков, а путем измерения изменений в голосе, обусловленных стрессом, который сопровождает ложные высказывания.

Суть работы анализаторов стресса в голосе объяснялась тем, что с помощью данных приборов якобы выделяются не воспринимаемые на слух акустические характеристики голоса, обусловленные стрессом. В частности, производители PSE утверждали, что этот прибор измеряет низкочастотную (около 10 Гц) модуляцию частоты основного тона голоса, обусловленную тремором мышц.

Проведенные экспертами-полиграфологами независимые исследования, исследования Американской ассоциации полиграфа (APA), а также тесты существующих на рынке приборов, проведенные Институтом полиграфа Минобороны США (DoDPI, показали, что точность этих приборов находится на уровне случайного угадывания [1].

Причину эксперты связывают с тем, что данные приборы лишь имитируют работу полиграфа. Они несравнимо беднее по возможностям измерения и анализа речевого сигнала классических систем распознавания речи, не учитывают опыт использования традиционного полиграфа с его обширной тактико-аналитической базой. В документации на эти приборы отсутствует описание принципов их работы под предлогом неразглашения коммерческой тайны, а инструкции применения отсутствуют или просто скопированы с инструкций некоторых моделей классического полиграфа.

Все это дискредитирует саму идею создания и использования речевого полиграфа на практике, требует постановки и решения задачи о применимости речевых анализаторов в оценке и интерпретации эмоциональных реакций человека по речи в части определения правдивости и искренности сказанного.

Для построения реально действующей системы распознавания эмоционального состояния говорящего по речи должны быть основательно проработаны измерительная база и алгоритмическая основа системы распознавания эмоций по речи.

Задача автоматического распознавания эмоциональной окрашенности звучащей речи является междисциплинарной и постоянно привлекает исследователей разных специальностей — не только лингвистов, но и математиков, программистов, психологов, физиологов [2].

Исследования ведутся по нескольким направлениям.

1. Модальность эмоций. Это традиционное направление работ психологов по изучению и классификации эмоций, выявлению эмоционально-смысловых коннотаций.
2. Нахождение объективных характеристик проявления эмоций в речи, связи эмоций с паралингвистическими, экстралингвистическими и артикуляционными особенностями речи. Одно из традиционных направлений работы лингвистов.
3. Способы извлечения эмоциональных характеристик из речевого сигнала. Построение пространства признаков для распознавания эмоций по речи. Работы ведутся смешанными коллективами, состоящими из физиологов, лингвистов, специалистов по автоматическому распознаванию речи.
4. Нахождение эффективных стратегий распознавания. Построение стратегий, алгоритмов и систем распознавания эмоций по речи. Верификация смыслов эмоциональных речевых реакций в зависимости от ситуативного контекста, выбор решений в отношении правдивости и искренности говорящего. Работы ведутся смешанными коллективами, состоящими из лингвистов, специалистов по автоматическому распознаванию речи, искусственному интеллекту.

По первому направлению можно выделить работы П. В. Симонова, К. В. Анохина, В. О. Леонтьева [3], К. Э. Изарда и A. R. Damasio, в которых были выделены группы эмоций. Среди этих групп принято выделять первичные и вторичные.

Первичные эмоции считаются базовыми, врожденными. Они включают в себя обобщенные, близкие к рефлексу («автоматические», или запрограммированные), страх и мгновенные реакции на стимулы, представляющие опасность. Они не предполагают сознательных размышлений и включают в себя шесть базовых эмоций, выделенных Дарвином: страх, гнев, отвращение, удивление, грусть и счастье; впрочем, по К. Изарду выделяют 11 фундаментальных (базовых) эмоций: радость, удивление, печаль, гнев, отвращение, презрение, горе-страдание, стыд, интерес-волнение, вина, смущение.

Вторичные эмоции — это более сложные эмоции, и они задействуют высшие центры коры головного мозга. Они могут заключать в себе базовые эмоции гнева или страха или

Рис. 1 Пример непрерывной шкалы эмоций

иметь более сложную структуру, например к ним добавятся сожаление, тоска, стыд, вина, зависть или ревность. Вторичные эмоции не являются автоматическими: они производятся мозгом, индивид думает о них и принимает решение, что с ними делать — какие действия лучше всего предпринять в той или иной ситуации.

Сознательные размышления и вторичные эмоции влияют на то, как индивид реагирует на ситуации, которые порождают первичные эмоции: он может отступить или смутиться предположив некоторую опасность, но, придя в себя, распознав и иначе оценив ситуацию, может, например, сделать вид, что ничего не случилось.

Главная проблема в обнаружении эмоционального состояния человека состоит в том, что все люди по-разному выражают свои эмоции. Кроме того, очень важно учитывать тонкие речевые компоненты и их изменение в процессе разговора. Поэтому исследователи от дискретной классификации эмоции и отнесения исследуемого фрагмента к какой-либо строго определенной категории эмоционального состояния переходят к описанию непрерывного эмоционального пространства, например к такому, как показано на рис. 1.

Преимущество такого подхода заключается в возможности выражать огромное количество эмоций: от «средней раздраженности» до «ярого гнева», — а также различать неуловимые отличия между очень схожими эмоциями.

Четырехмерную сферическую модель эмоций предложила группа исследователей в публикации В. А. Вартанова [4]. Построение модели проводилось экспериментально с помощью многомерного шкалирования субъективных различий между эмоциональными состояниями, задаваемыми специально созданными образцами. Чтобы уровнять и сделать определенным содержание этих образцов, в эксперименте использовалось одно и то же слово, произнесенное в разных эмоциональных состояниях. В одной серии использовалось слово «да», а в другой — «нет». Полученные параметры (факторы) характеризовались как бимодальные спектральные фильтры. Из них выделили четыре измерения: лучше–хуже, удивление–уверенность, симпатия–равнодушие, активное–пассивное отвержение.

По второму направлению лингвисты и психологи выявляют эмоциональные составляющие речи, анализируя ее паралингвистические, экстралингвистические и артикуляционные особенности.

Из прикладной лингвистики и апеллятивной фонетики известно, что многие признаки эмоционального состояния, искренности и правдивости говорящего содержатся в мело-

дике, акцентуации, смене темпа и ритма речи, особенностях артикулирования, дрожании голоса — особенно в стрессовых ситуациях, например при ответах собеседника на неожиданные «неудобные» вопросы.

Например, некоторыми исследователями установлена связь направления движения высоты голоса с положительными или отрицательными эмоциями: понижение высоты — с приятными эмоциями, а ее повышение соотносят с удивлением или страхом. Большое значение придают специалисты завершающему фрагменту мелодического контура фразы, поскольку он может информировать не только о повествовательном, вопросительном или восклицательном типе предложения, но и об отношении говорящего к теме высказывания, ситуации общения, к собеседнику.

Эмоциональную составляющую речитации обнаруживают не только в просодии (мелодике, ритме, акцентуации, темповой и ритмической динамике речи), но и в артикулировании (характере произнесения гласных, согласных, слогов, слов). Данные апеллятивной фонетики содержат важную информацию об эмоциональном состояния говорящего, что позволяет использовать артикуляционные модели в задачах распознавания эмоций по речи [5]. Экстралингвистические особенности речи проявляются в дрожании голоса, паузах, придыхании, заикании, покашливании, смехе [6]. Перед исследователями стоит задача установить их эмоциональную обусловленность, соотнеся с описанной выше модальностью эмоций.

По третьему направлению — основная задача получения признаков эмоциональной составляющей речи состоит в том, чтобы преобразовать звуковую волну в такое признаковое пространство, в котором множество объектов одного класса будет сгруппировано вместе, а множество объектов альтернативных классов максимально разнесено. Из всего спектра работ на современном этапе можно выделить четыре группы объективных признаков и соответствующих методов, позволяющих различать речевые образцы: спектрально-временны́е, кепстральные, амплитудно-частотные и признаки на основе нелинейной динамики [4–9].

Четвертое направление — нахождение эффективных механизмов и стратегий распознавания для создания речевого полиграфа; построение алгоритмов, сценариев и, наконец, систем распознавания правдивости и искренности говорящего по речи [10]; верификация смыслов эмоциональных речевых реакций в зависимости от ситуативного контекста; выбор решений.

Нельзя не признать, что одни и те же феномены эмоциональной речи в зависимости от ситуации могут быть интерпретированы по-разному. При маркировке тех моментов в высказывании, где проявляется волнение, в контексте ситуации может быть учтено влияние обстоятельств на общую картину эмоций и смысл происходящего в момент речевого высказывания [11].

Например, в зависимости от ситуации нескрываемый гнев говорящего (проявляется, например, в характерной смене темпа и ритма речи, тщательном выговаривании согласных в словах) может свидетельствовать о неверно высказанном предположении, содержащемуся в вопросе, заданном испытуемому, или о его отношении к самой ситуации допроса, к допрошивающему; растерянность и смущение, проявляющиеся в неуверенной речи, могут говорить как о страхе разоблачения, так и о непонимании вопрооса. Дрожание голоса в зависимости от ситуации может свидетельствовать об обиде, страхе, гневе или, наоборот, радости.

Для сопоставления речевых реакций с ситуацией, в которых они проявлялись, нами был разработан нормативный язык описания модели ситуаций, опирающийся на наше понимание их морфологии [11].

Настоящая работа посвящена экспериментам по созданию системы распознавания правдивости и искренности говорящего.

2 Постановка задачи

Перед началом эксперимента по автоматическому или экспертному распознаванию правдивости и искренности говорящего по речи исследователи в качестве исходных данных, как правило, располагают:

- речевым сигналом, представленным в виде дискретной функции от времени — последовательности временны́х отсчетов $C(k)$, где k — номера отсчетов сигнала по оси времени с фиксированным шагом;
- информацией о характере говорящего, т. е. о характерных для него когнитивных, регулятивных и коммуникативных особенностях проявления эмоций E;
- первичной характеристикой расследуемой ситуации в виде совокупности обстоятельств, описанных некоторым нормативным языком G.

Перед нами стоит задача выявления эмоциональной составляющей речи по паралингвистическим, экстралингвистическим, артикуляционным особенностям высказывания и распознавание смысла эмоции индивида в контексте ситуации дознания. Эти эмоции, являясь непроизвольной реакцией индивида на попытку исследователя тем или иным образом прояснить исследуемую ситуацию, должны были бы позволить сделать заключения в отношении сказанного испытуемым и прояснить как позицию испытуемого в ситуации дознания, так и общую картину происшествия.

Однако неоднозначность эмоционально-смысловых коннотаций в проекции на реконструируемую картину происшествия может привести к существенным ошибкам и делает необходимой выработку специальной стратегии для выбора решений в распознавании смысла речевых эмоций.

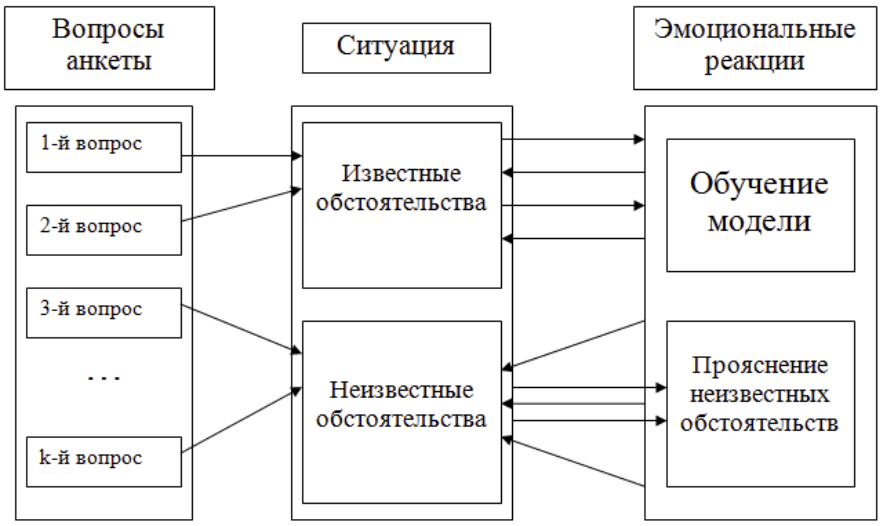


Рис. 2 Обучение модели и распознавание эмоциональной реакции человека в ситуативном контексте

В данной работе мы исходим из предположения о том, что верификация смыслов эмоциональных проявлений в речи по выработанным признакам становится возможной с помощью сопоставления эмоционально-смысловых коннотаций с ситуативным контекстом при условии применения специальных процедур опроса испытуемого в процессе реконструкции расследуемого события (рис. 2).

3 Описание эксперимента

В настоящей работе распознавание эмоций и заключение о правдивости и искренности испытуемого опиралось на:

- паралингвистические особенности речи (т. е. ее мелодику, акцентуацию, темпоритм, см. рис. 3–5), характерные для индивида;
- индивидуальные особенности артикулирования;
- экстралингвистические особенности высказывания; к ним относятся — паузы, смех, покашливание, вздохи, плач, мычание, заикание, дрожание голоса;
- знания об эмоционально-смысловых коннотациях, характерных для речи испытуемого;
- соотнесение эмоциональности высказывания с ситуативным контекстом.

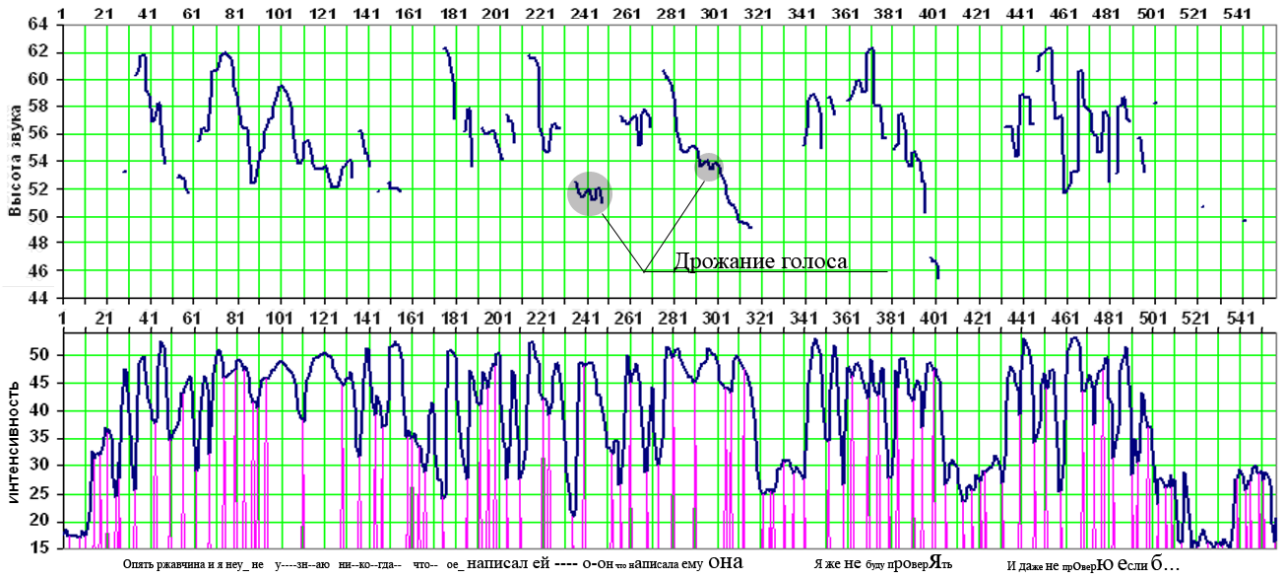


Рис. 3 Первичная сегментация по минимумам интенсивности звука в высказывании «что написал...» и характерные фрагменты дрожания голоса в словах «он», «она»

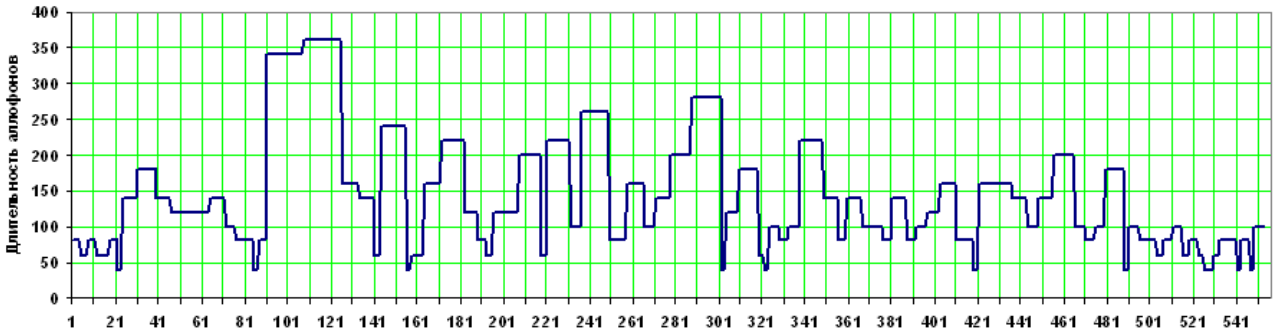


Рис. 4 Положение длительностей аллофонов в высказывании «Что написал»

Рис. 5 Динамика темпа произнесения гласных

3.1 Измерительная база

На первом этапе обработки данных речевой сигнал $C(k)$ подвергался спектральному анализу посредством быстрого преобразования Фурье (БПФ) с последовательно сдвигаемым взвешенным окном. Вычислялся динамический спектр в виде последовательности значений кратковременных энергетических спектров $S(w, i)$, измеренных в моменты времени каждые 20 мс, траектории максимумов трех первых формант $F(j, i)$, кривые интенсивности в низком, среднем и высоком частотных диапазонах $F(l, i)$ — так называемая «гребенка», амплитудная огибающая общей интенсивности $A(i)$ и звуковысотный контур речевой просодии $P(i)$, рассчитав для этого по специальным алгоритмам траекторию основного тона.

На основании этих данных:

- произвели первичную сегментацию по минимумам интенсивности звука (см. рис. 3);
- произвели маркировку аллофонов $A(m)$; сегменты идентифицировали по их спектральных характеристикам и на основании справочных материалов или экспертных оценок по типу аллофон гласного/согласного (вокализованный, щелевой, взрывной и т. п.);
- скорректировали, перегруппировали первичную сегментацию и вычислили длительности звуков, соответствующих гласным и согласным (рис. 6);

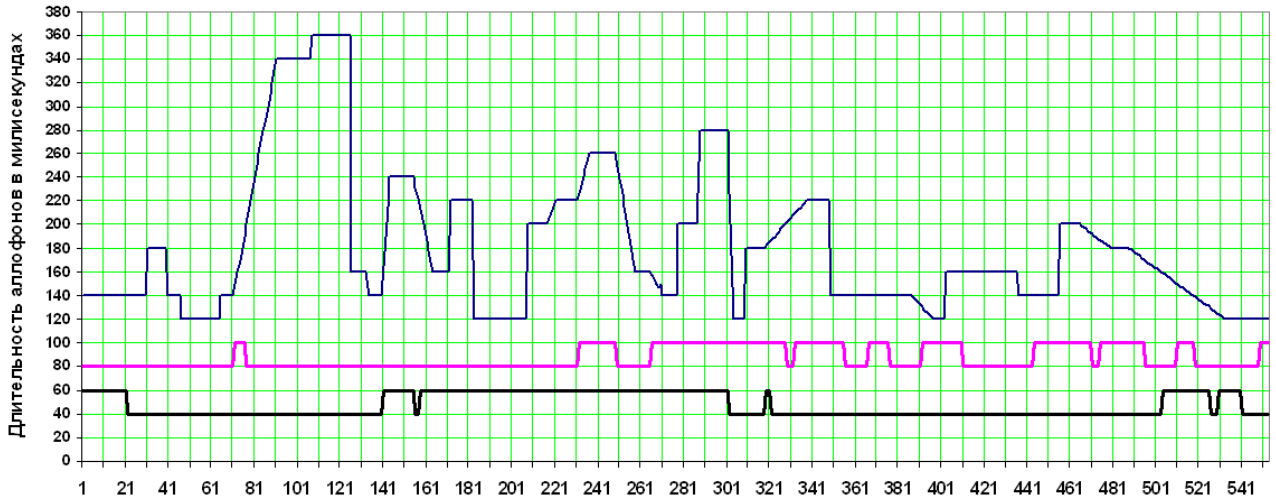


Рис. 6 Графики динамики длительностей гласных (верхний), сонорных, щелевых (средний) и взрывных (нижний) согласных в высказывании «что написал...»

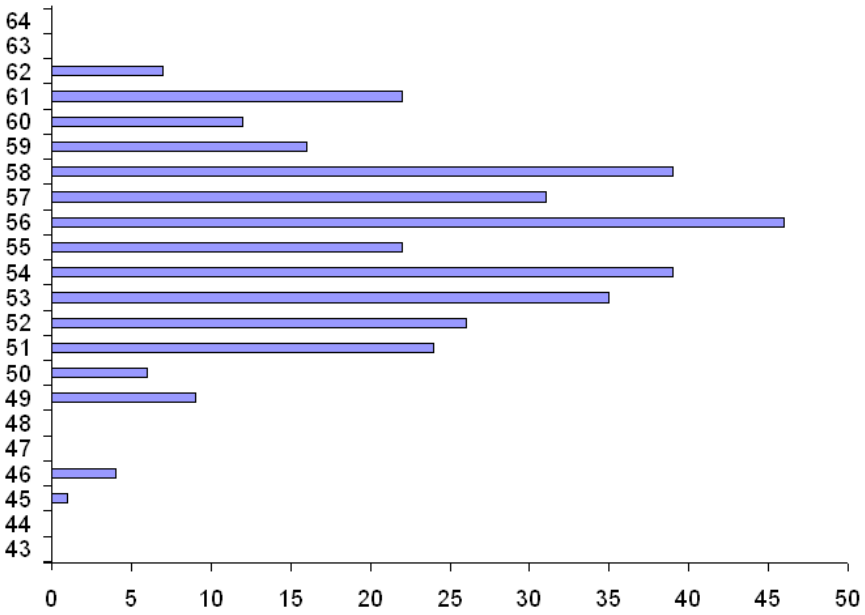


Рис. 7 Гистограмма высоты голоса в высказывании «что написал...»

– выявили просодию, т. е. паралингвистические особенности речи, как то:
  - высоту голоса привели к непрерывной музыкальной шкале стандарта MIDI, где нота «Do» первой октавы соответствует 52-м, «Re» — 54-м и т. д., по этой шкале анализировали мелодику речевого высказывания;
  - динамический темп речи (см. рис. 5);
– распознали ритмические формы;
– определили особенности интонирования, например установили присутствие элементов контрастно-регистрового интонирования, что наглядно видно на рис. 3 и 7 во временно́м диапазоне 273–321 отсчетов: присутствует бросок высоты голоса на октаву вниз менее чем за 1,5 с.

Полученные данные использовались:

– на этапе обучения модели при экспертной разметке на эпизоды, свидетельствующие об эмоциональности речи и выявлении характерных для индивида эмоционально-смысловых коннотаций;
– на этапе распознавания для выявления эмоциональной составляющей речи, заключения об искренности говорящего и правдивости сказанного.

3.2 Выявление признаков эмоций в речи

В результате серии опытов и экспертных заключений об эмоциональности речевых фрагментов наиболее информативными оказались следующие признаки:

– длительность ударных и безударных гласных по отношению к средней длительности их произнесения в текущем эпизоде позволяют выявить фразовые и эмоциональные акценты;
– удлинение предударных щелевых или сонорных согласных является средством эмоционального усиления акцента говорящим;
– преувеличенная акцентуация ударного слога в слове за счет интенсивности звука голоса — свидетельство эмоционального возбуждения;

– акцентуация за счет увеличения длительности ударных гласных и предударных согласных — свидетельство желания убедить собеседника;
– обратная величина динамики длительности гласных дает динамику темпа речи (см. рис. 5);
– изменение темпа речи на уровне слова, фразы, высказывания свидетельствует об осмысленном проявлении отношения говорящего к смыслу высказывания, желании выделить или скрыть это отношение;
– эмоциональное усиление акцента в слове, фразе часто сопровождается «двойной акцентуацией» внутри ударных гласных, появляются два максимума на кривой интенсивности, которые состоят из двух сегментов первичной разбивки;
– речевой ритм, построенный на соотношении длительностей соседних ударных и безударных гласных, помимо того что позволяет распознавать акцентуацию, уточнять наличие структурирующих (словесных, фразовых) и эмоциональных акцентов, выявляет мультипликативные формы (например, скандирование), чаще всего имеющие эмоциональную природу;

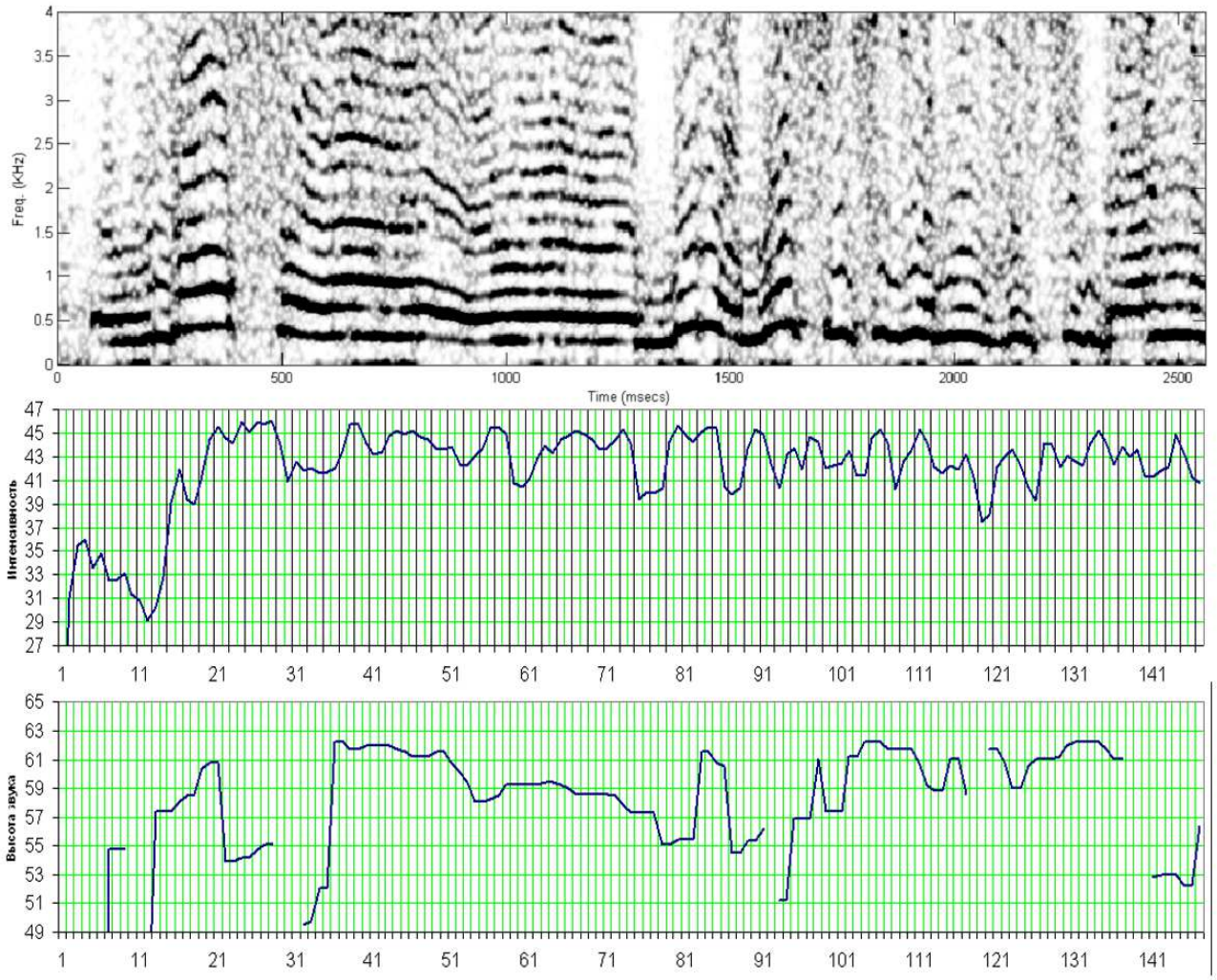


Рис. 8 Интенсивность и высота звука в высказывании «У нас э-э-э-э было восемь судебных заседаний»

– дрожание голоса (относится к экстралингвистическим элементам речи, см. рис. 3) свидетельствует о непроизвольно проявляющемся волнении — чаще всего от негодования, страха, обиды или, наоборот, от радости, восторга; и если негодование, радость и восторг обычно сопровождаются повышенным уровнем интенсивности звука, то страх и обида проявляются средним или пониженным уровнем интенсивности;
– характерное периодическое чередование взрывных участков и пауз (смычек) свидетельствует о смехе, покашливании в речи;
– длительные (порядка секунды и более) вокализованные «А», «Э», «М» свидетельствуют о неуверенной речи, неподготовленности речевого высказывания.

На рис. 8 изображен фрагмент амплитудной и высотной характеристик типичного неуверенного «блеяния» другого диктора во фразе «У нас э-э-э-э было восемь судебных заседаний».

На основании предварительных данных о проявлении эмоций в речи в виде набора эмоциональных речевых признаков $\boldsymbol{me}$ мы построили, а впоследствии верифицировали с помощью экспертных оценок характерный для испытуемого индивида набор эмоционально-смысловых коннотаций.

В результате наших исследований [5, 6, 8] были выявлены ранее неизученные связи речевых признаков с модальностью эмоций, такие как:

– контрастно-регистровое интонирование, означающее испуг, панику (см., например, октавный бросок в траектории высоты голоса в слове «она» на рис. 3);
– смена ритма со сложного на простой, означающая раздражение, гнев;
– двойная акцентуация гласных, означающая возмущение;
– подмена гласных в акцентируемом слоге, например «а» на «ы» (пример во фразе «сама понимаешь» первая гласная «а» звучит как «ы»), свидетельствует об агрессивности, гневе, злости, возмущении; вычисляется по артикуляционным моделям гласных на основании значений первых трех формант в распознаваемом сегменте (рис. 9–11).

Кроме того, были приняты во внимание взаимозависимости высоты голоса, интенсивности звука, темпа, разборчивости и уверенности речи, полученные другими исследователями [4, 7, 9, 12]:

– явно высокий звук — энтузиазм, радость, испытуемый заинтересован и проявляет интерес;

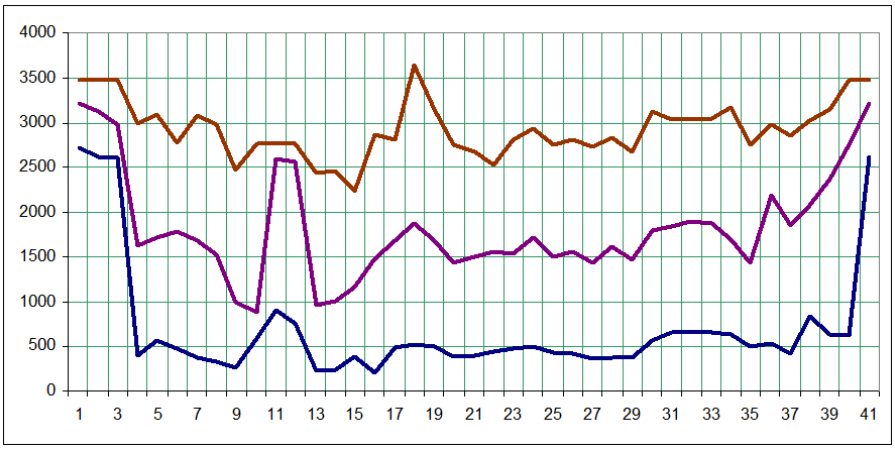


Рис. 9 Траектория трех первых формант во фразе «сама понимаешь»

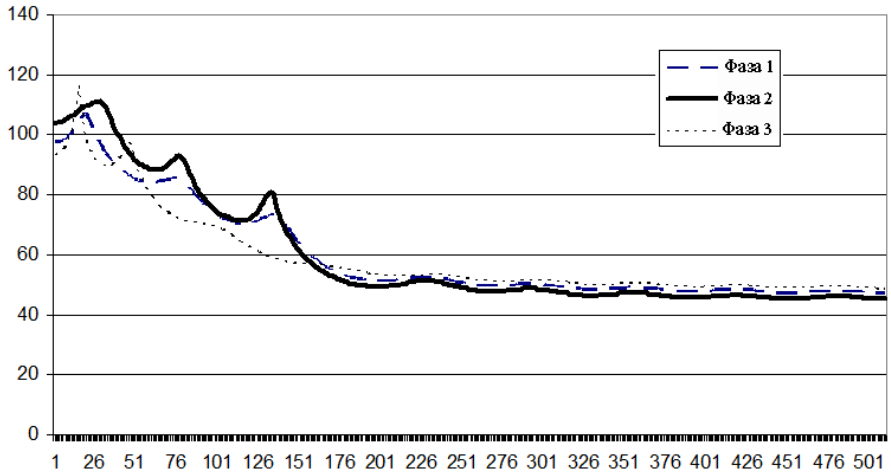


Рис. 10 Фазы LPC-огибающей динамики спектра произнесения гласной в первом слоге слова «сама»

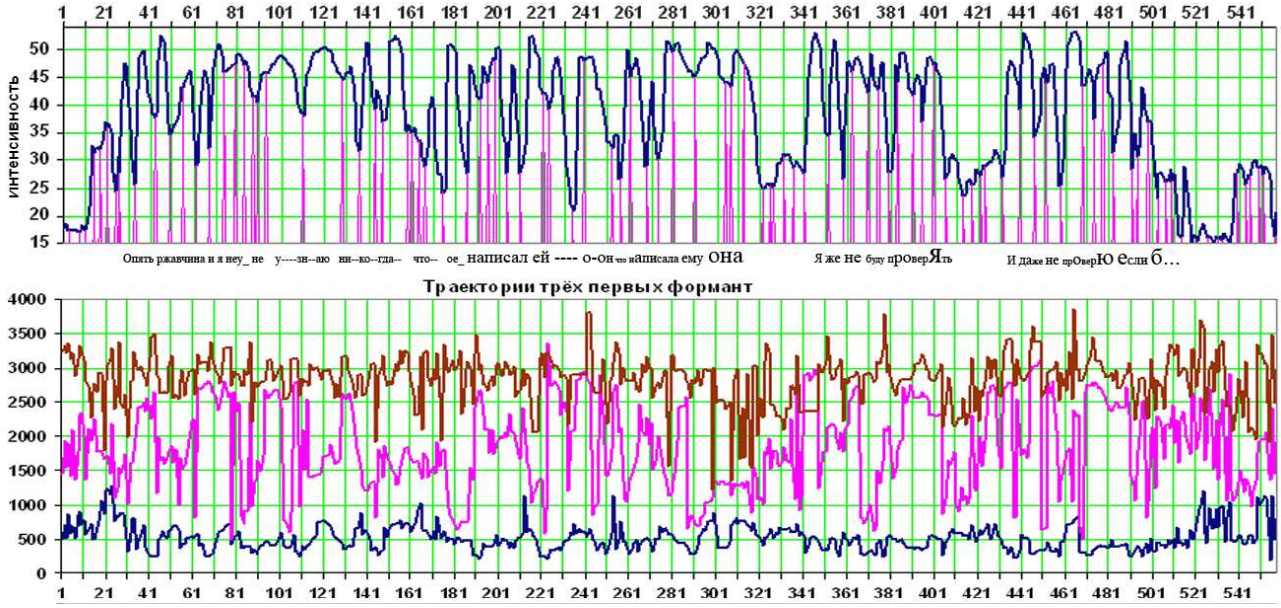


Рис. 11 Графики интенсивности и траектории трех первых формант для выявления характера произнесения гласных

- чрезмерно высокий, пронзительный — беспокойство;
- мягкий и приглушенный, с понижением интонации к концу каждой фразы — печаль, усталость;
- форсирование звука — напряжение, обман;
- быстрая речь — очевидная взволнованность — желание убедить или уговорить кого-то;
- медленная речь — высокомерие, усталость, угнетенное состояние;
- прерывистая речь – неуверенность;
- лаконичность и решительность речи — явная уверенность;
- заикание — напряженность или обман;
- нерешительность в подборе слов — неуверенность в себе или намерение внезапно удивить чем-то;

– появление речевых недостатков (повторение или искажение слов, обрывание фраз на полуслове) — несомненное волнение, но иной раз и желание обмануть;
– опускание речевых пауз — напряжение;
– слишком удлиненные паузы — незаинтересованность или несогласие.

Для создания модели реакций испытуемого был выделен специальный этап исследования — создание массива данных, где накапливались данные о речевых реакциях испытуемого, проводилось выделение значимых параметров.

3.3 Построение и обучение модели

Для разбиения текущих значений речевых параметров на классы (группы признаков) таких непрерывных параметров, как мелодический контур, огибающая интенсивности речевого сигнала или динамика темпа речи, длительность пауз, использовался метод кластеризации, вероятностный подход. Предполагалось, что каждый рассматриваемый на этапе обучения объект (эмоциональная речевая реакция) относится к одному из k классов обучающих выборок. Для определения центроида кластера вычислялась медиана и производилось обучении модели.

Принадлежность сегмента определялась после вычисления его метрики по группе указанных выше параметров в соотнесении с алфавитом сегментов, выявленных и маркированных в процессе экспертной оценки на этапе обучения модели; его многомерная классификация и соответственно маркировка осуществлялись при выполнении ряда условий.

Обозначим множество темпорально-акустических характеристик речевого высказывания как $\boldsymbol{M}$ (из которых подмножество $\mathbf{me}$ свидетельствует об эмоциональной окраске, так что $\mathbf{me}(\boldsymbol{i})$ — набор признаков эмоций в речи, где $i = 1, \ldots, N$, и N — количество классов эмоциональной окраски, отражающихся в речевых параметрах), совокупность эмоционально-смысловых коннотаций индивида как $\boldsymbol{E}$, морфологию ситуаций дознания и реконструируемого происшествия как упорядоченные множества $\boldsymbol{G_1}$ и $\boldsymbol{G_2}$.

В данной работе была поставлена задача выбора решений об искренности говорящего и правдивости сказанного им при распознавании эмоций по речи в связанной системе $\boldsymbol{M}, \boldsymbol{E}, \boldsymbol{G_1}, \boldsymbol{G_2}$. Перед проведением эксперимента была построена модель возможных эмоциональных реакций испытуемого $\boldsymbol{G_1}$–$\boldsymbol{G_2}$–$\boldsymbol{M}$–$\boldsymbol{E}$, которая во время эксперимента обучалась по примерной схеме, отображенной на рис. 2.

Итак, у нас есть упорядоченное подмножество $\mathbf{me}$ множества $\boldsymbol{M}$ признаков, распознанных экспертами как характеристики, которые свидетельствуют о волнении говорящего или желании диктора выделить слово, артикулируя звуки в нем особым образом. У нас это множество разбито на N классов и представлено массивом данных $\mathbf{me}(k, i, m, L)$, где каждый k-й элемент из $\boldsymbol{M}$ отнесен к i-му классу и каждому классу поставлено в соответствие значения из $\boldsymbol{S}$ — группы частично упорядоченных параметров, представленных массивом $\boldsymbol{S}(i, m, c, d)$, где i — имя (номер) класса; m — имя (номер) параметра; c — положение центроида класса; d — медиана класса i.

Тогда $\boldsymbol{\Delta\widehat{L}}$ — ближайшее расстояние между кластерами значений (векторная разность) параметров i-х классов массива $\boldsymbol{S}$ и соответствующими значениями параметров распознаваемого $(n+1)$-го сегмента в пространстве признаков, т. е.

$$\Delta\widehat{L} = \arg\min_i [\widehat{M}(k+1) - \widehat{S}(i)].$$

Эмоциональная окраска сегмента может рассматриваться как вероятность, вытекающая из величины отклонения его параметров от некоторых «нормальных» значений

для данного контекста. Здесь мы опираемся на имеющиеся в наличии аналогичные по контексту артикуляционные позиции, которые могут быть распознаны как сегменты, соответствующие «спокойному артикулированию».

Для фиксации наличия признаков эмоций в речи, таких как смена ритма со сложного на простой (которая устанавливалась с помощью автокорреляционной функции длительностей гласных), контрастно-регистровое интонирование (которое определялось на плоскости по расстоянию между пиками функции плотности вероятности высоты голоса и временно́й дистанции между их значениями в речевом фрагменте) и т. п. выявлялся сам факт наличия такого признака, т. е. использовалась бинарная оппозиция — есть, нет (true, false).

3.4 Пример выбора решения

В выборе решения о подлинном смысле и правдивости сказанного учитывалось соотнесение трёх групп признаков:

- акустико-временны́х — тональных, спектродинамических и темпоральных характеристик речи и вычисленных по ним данных о просодии и артикуляции высказывания;
- эмоционально-смысловых коннотаций речи;
- ситуативных — данных о ситуативном контексте высказываний; при этом рассматривалась морфология двух разных, но связанных между собой ситуаций — текущая ситуация дознания и модель реконструируемой следствием цепи событий.

Здесь вопросы дознания к испытуемому на этапе обучения модели наряду с дискретными состояниями описания ситуации дознания, которые относятся ко множеству обстоятельств $\boldsymbol{G_1}$, задаются по известным обстоятельствам как $\boldsymbol{G_1}$, так и $\boldsymbol{G_2}$. По реакции индивида на вопросы по заранее известным обстоятельствам происходит обучение модели распознавания.

При накоплении достаточной представительности обучаемой модели испытуемому задаются вопросы в отношении неизвестных следствию обстоятельств $\boldsymbol{G_2}$, эмоциональные речевые реакции $\mathbf{me}$ испытуемого на вопросы из множества $\boldsymbol{G_2}$ соотносятся со множеством смысловых коннотаций $\boldsymbol{E}$, на основании чего делается вывод об искренности и правдивости ответа. При этом информативными оказываются как искренние, правдивые ответы, так и ложные, так как они свидетельствуют о попытке сокрытия обстоятельств, которые нужны для дополнения описания $\boldsymbol{G_2}$. В этом случае в части модели ситуации дознания $\boldsymbol{G_1}$ может быть сформирован дополнительный сценарий для прояснения обстоятельств, которые пытался скрыть испытуемый.

Анализируя исходное речевое высказывание $C(k)$, касающееся описания ситуции $\boldsymbol{G_2}$ в ряду темпорально-акустических признаков $\boldsymbol{M}$, мы выделяем из перечисленных 28 признаков 8 значимых для данного высказывания и устанавливаем эмоционально-смысловые коннотации $\boldsymbol{E}$, связанные с элементами исследуемой ситуации $\boldsymbol{G_2}$, одновременно выявляя неоднозначности эмоционально-смысловых коннотаций.

Так, например, увеличение темпа речи во временно́м диапазоне 125–210 отсчетов на рис. 3 и снижение темпа в диапазоне 210-250 отсчетов может свидетельствовать как о желании убедить собеседника, так и о неуверенности говорящего, его волнении.

Природа волнения становится понятной из сопоставления с другими эмоциональными признаками в этом же речевом фрагменте — из дрожания голоса на указанном участке снижения темпа на слове «он» и следующим за ним фрагменте (273–321 отсчетов) с эмоциональными признаками — контрастно-регистровом интонирования и дрожании голоса на

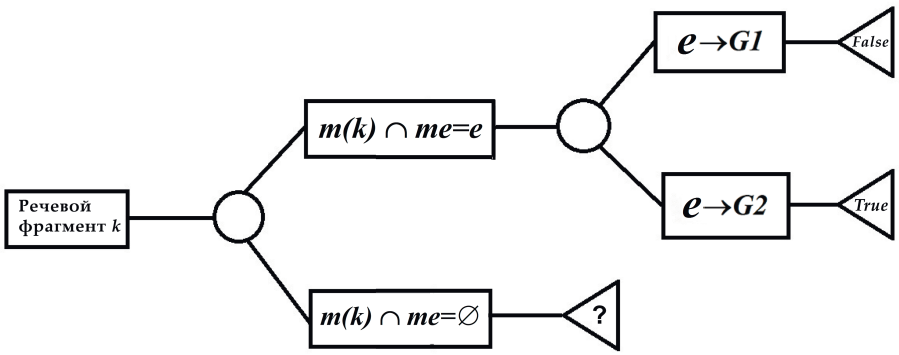


Рис. 12 Схема вычисления целевой переменной на основании акустико-темпоральных и ситуативных признаков

слове «она» с двойной акцентуацией на последнем слоге этого слова, свидетельствующими о страхе и глубокой обиде.

Ситуативный контекст проясняет смысл эмоционального всплеска — здесь речь идет о подозрении в супружеской неверности. Функция испытуемого — пострадавший.

Соотнесение многофакторных признаков и заключение об искренности и правдивости говорящего осуществлялось с помощью дерева принятия решений. Структура дерева представляла собой «листья» и «ветки». На ребрах («ветках») дерева решения записывались атрибуты (речевые признаки $\boldsymbol{M}$ и ситуативный контекст $\boldsymbol{G_1, G_2}$), от которых зависела целевая функция (смысл эмоциональной реакции), в «листьях» были записаны значения целевой функции, а в остальных узлах — атрибуты, по которым разветвляется дерево и из которых одна из ветвей приводит к конечному значению целевой функции.

Таким образом, для классификации очередного фрагмента речевой реакции нужно спуститься по дереву до листа и выдать соответствующее значение.

Эта схема отражает модель, которая вычисляет значение целевой переменной на основе нескольких акустико-темпоральных и ситуативных переменных параметров (признаков) на входе (рис. 12).

По указанной схеме в процессе эксперимента был реализован алгоритм распознавания в связанной системе $\boldsymbol{M, E, G_1, G_2}$ — от последовательности эмоционально значимых речевых фрагментов и алфавита эмоционально-смысловых коннотаций до процессов дознания и реконструкции исследуемой ситуации.

4 Заключение

В работе описан опыт выбора решений в системе распознавания эмоционального состояния человека по речи. Анализ эмоциональных проявлений на основе соотнесения пара- и экстралингвистических особенностей и артикуляционных моделей речи с их эмоционально-смысловыми коннотациями показывает неоднозначность этих коннотаций, что в проекции на реконструируемую картину происшествия и ситуацию дознания может привести к существенным ошибкам распознавания. Предложена стратегия выбора решений распознавания эмоционального состояния человека по речи в связанной системе темпорально-акустических, эмоционально-смысловых и ситуационных зависимостей. При настоящем подходе верификация смыслов эмоциональных речевых реакций становится возможной благодаря именно сопоставлению с ситуативным контекстом.

Литература

[1] *Князев В., Варламов Г.* Полиграф и его практическое применение. — Принт-Центр, 2012. 859 с.

[2] *Кальян В. П.* Музыка, речь и компьютер. — М.: ВЦ РАН, 1998. 38 с.

[3] *Леонтьев В. О.* Десять нерешенных проблем теории сознания и эмоций. — Одесса, 2008. http://polatulet.narod.ru/dvc/com/vleontiev_problems.html.

[4] *Вартанов А. В.* Антропоморфный метод распознавания эмоций в звучащей речи // Национальный психологический ж., 2013. № 2[10]. С. 69–79. http://www.psy.msu.ru/science/npj/journals/npj-no10-2013.pdf.

[5] *Кальян В. П.* Исследование применимости артикуляционных моделей в задачах распознавания эмоций по речи // Докл. 9-й Междунар. конф. «Интеллектуализация обработки информации». — М.: ТОРУС ПРЕСС, 2011. С. 334–349.

[6] *Кальян В. П.* Построение алгоритмов распознавания эмоционального состояния человека по пара и экстралингвистическим особенностям речи // Модели и методы распознавания речи. — М.: ВЦ РАН им. А. А. Дородницына, 2010. С. 24–46.

[7] *Schuller B., Steidl S., Batliner A.* The INTERSPEECH 2009 emotion challenge // Interspeech, 2009. Т. 2009. С. 312–315. http://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_0312.pdf.

[8] *Кальян В. П.* Разработка алгоритмов распознавания эмоционального состояния человека по паралингвистическим особенностям речи // Докл. 15-й Всеросс. конф. «Математические методы распознавания образов». — М.: МАКС-Пресс, 2011. С. 334–349.

[9] *Брестер К. Ю.* Коллективный эволюционный метод многокритериальной оптимизации в задачах анализа речевых сигналов. Дисс. ... канд. техн. наук. — Красноярск: 2013. 143 с. http://research.sfu-kras.ru/sites/research.sfu-kras.ru/files/Dissertaciya_Brester_K.Yu_.pdf.

[10] *Кальян В. П.* Архитектура системы распознавания эмоционального состояния человека по речи // Модели и методы распознавания речи. — М.: ВЦ РАН им. А. А. Дородницына, 2013. С. 89–98.

[11] *Кальян В. П.* Морфология ситуации в системе распознавания эмоционального состояния человека по речи // Модели и методы распознавания речи. — М.: ВЦ РАН им. А. А. Дородницына, 2012. С. 92–102.

[12] *Сидоров К. В., Филатова Н. Н.* Анализ признаков эмоционально окрашенной речи // Известия ЮФУ. Технические науки, 2012. Т. 134. № 9. С. 39–45. http://eprints.tstu.tver.ru/69/1/3.pdf.

Поступила в редакцию 13.09.2016

Decision support in process of recognizing emotion in speech

V. P. Kalyan

vkalyan@mail.ru

Federal Research Center "Computer Science and Control" of RAS

44/2 Vavilova Str., Moscow, Russia

Background: Commercial devices, presenting themselves as "analyzers of stress in a voice," have been appearing in the market for lie detection services for more than 40 years. Those devices, unlike polygraphs, were claimed to be capable of establishing insincerity without requiring any connection to human body via sensors, but by measuring changes in one's voice

caused by raised stress level that is provided by making false statements. The independent researches of the devices existing in the market, conducted by polygraphology experts, American Association of a Polygraph (MACAW), Institute of Polygraph Tests of the USA Ministry of Defence (DoDPI), proved that the accuracy of those devices drops to the level of random guessing.

Methods: This work describes the experience of decision making in the system designed to recognize the emotional state of a person by his/her speech, concerning truthfulness and sincerity of what is being said. The information value of the recognition-measuring base is analyzed on the basis of paralinguistic, articulation, and extralinguistic speech features, regarding also individual emotional and semantic connotations of the testee's speech and the algorithms helping to recognize emotions by speech are described. We make the choice from a number of decisions and verify them regarding to the speaker‘s sincerity and truthfulness and concerning situational context as well.

Results: The analysis of emotional expressions based on matching para- and extralinguistic features and articulatory model of speech to their emotional and semantic connotations shows certain ambiguity of those connotations. It can lead to serious essential mistakes while recognizing if projected to the reconstructed accident picture and a situation of inquiry. A strategy for choosing decisions for identifying one's emotional state by his speech is proposed within within a related system of temporal and acoustic, emotional and semantic and situational dependences. This way gives one an opportunity to verify the meanings of emotional speech reactions due to correlating with the situational context.

Keywords: *recognition of emotions; emotional speech; decision-making tree; space of speech signs; paralinguistic features of speech; articulation models; spectral dynamics; speech formant; sound pitch; sound altitude*

DOI: 10.21469/22233792.2.4.07

References

[1] Kniazev, V., and G. Warlamov. 2012. *Poligraph i ego prakticheskoe primenenie* [Polygraph and its practical application]. Print-Center. 859 p.

[2] Kalyan, V. P. 1998. *Musyka, rech' i komp'yuter* [Music, speech, and computer]. Moscow: A. A. Dorodnitsyn CC RAN. 38 p.

[3] Leontiev, V. O. 2008. *Desyat' nereshennykh problem teorii soznaniya i emotsiy* [Ten unsolved problems in the theory of consciousness and emotions]. Odessa. Available at: `http://polatulet.narod.ru/dvc/com/vleontiev_problems.html` (accessed April 7, 2017).

[4] Vartanov, A. V. 2013. Antropomorfnyy metod raspoznavaniya emotsiy v zvuchashchey rechi [Anthropomorphic method of emotion recognition in sounding speech]. *Natsyonalnyi psikhologitcheskiy zh.* [National Psychological J.] 2(10):69–79. Available at: `http://www.psy.msu.ru/science/npj/journals/npj-no10-2013.pdf` (accessed April 7, 2017).

[5] Kalyan, V. P. 2012. Issledovanie primenimosti artikulyatsionnykh modeley v zadachakh raspoznavaniya emotsiy po rechi [Study of the applicability of articulatory models in speech recognition problems by speech]. *9th Conference (International) on Intellectualization of Information Processing Proceedings*. Moscow: TORUS PRESS. 498–502.

[6] Kalyan V. P. 2010. Postroenie algoritmov raspoznavaniya emotsional'nogo sostoyaniya cheloveka po para i ekstralingvisticheskim osobennostyam rechi [The construction of algorithms for recognizing the emotional state of a person according to para- and extralinguistic features of speech]. *Modeli i metody raspoznavaniya rechi* [Models and methods of speech recognition]. Moscow: A. A. Dorodnitsyn CC RAS. 24–46.

[7] Shuller, B., S. Steidl, and A. Batliner. 2009. The INTERSPEECH 2009 emotion challenge. *Interspeech* 2009:312–315. Available at: http://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_0312.pdf (accessed April 7, 2017).

[8] Kalyan, V. P. 2011. Razrabotka algoritmov raspoznavaniya emotsional'nogo sostoyaniya cheloveka po paralingvisticheskim osobennostyam rechi [Development of algorithms for recognizing a person's emotional state by paralinguistic features of speech]. *Dokl. 15-y Vseross. konf. "Matematicheskie metody raspoznavaniya obrazov"* [15th All-Russian Conference on Mathematical Methods of Sspeech Recognition Proceedings] Moscow: MAKS-Press. 334–349.

[9] Brester, K. J. 2013. Kollektivnyy evolyutsionnyy metod mnogokriterial'noy optimizatsii v zadachakh analiza rechevykh signalov [Collective evolutionary method of multicriteria optimization in problems of analysis of speech signals]. PhD Diss. Krasnoyarsk. 143 p. Available at: http://research.sfu-kras.ru/sites/research.sfu-kras.ru/files/Dissertaciya_Brester_K.Yu_.pdf (accessed April 7, 2017).

[10] Kalyan, V. P. 2013. Arkhitektura sistemy raspoznavaniya emotsional'nogo sostoyaniya cheloveka po rechi [Architecture of the system of recognition of the emotional state of a person by speech]. *Modeli i metody raspoznavaniya rechi* [Models and methods of speech recognition]. Moscow: A. A. Dorodnitsyn CC RAS. 89–98.

[11] Kalyan, V. P. 2012. Morfologiya situatsii v systeme raspoznavaniya emotsional'nogo sostoyaniya cheloveka po rechi [Morphology of the situation in the system of recognition of the emotional state of a person by speech]. *Modeli i metody raspoznavaniya rechi* [Models and methods of speech recognition]. Moscow: A. A. Dorodnitsyn CC RAS. 92–102.

[12] Sidorov, K. V., and N. N. Filatova. 2012. Analiz priznakov emotsional'no okrashennoy rechi. *Izvestija UFU* 134(9):39–45. Available at: http://eprints.tstu.tver.ru/69/1/3.pdf (accessed April 7, 2017).

Received September 13, 2016