Автоматическое распознавание эмоций по речи с использованием метода опорных векторов и критерия джина

Хитров М.В., *кандидат технических наук,*

Давыдов А.Г., *кандидат физико-математических наук,*

Киселёв В.В., *кандидат технических наук,*

Ромашкин Ю.Н., *кандидат технических наук,*

Ткаченя А.В., *младший научный сотрудник*

В статье излагаются результаты исследования эффективности автоматического распознавания базовых эмоций в русской речи. Исследование включает выбор наиболее эффективных информативных признаков речи, построение мультиклассового классификатора и эксперименты с использованием собранного корпуса эмоциональной речи.

• автоматическое распознавание эмоций • русский язык • SVM-классификатор

The results of investigation of the effectiveness of the auto-detected-of basic emotions in the Russian language are described. The study includes a selection of the most efficient informative speech parameters, multi-class classifier building and experiments with the use of the assembled emotion speech base.

• automatic recognition of emotions • Russian language • SVM classifier

Введение

Исследование паралингвистических средств речевой коммуникации включает определение довольно разнообразных характеристик: психоэмоционального состояния человека, его пола и возраста, стиля разговора, уровня заинтересованности и даже наличия алкогольной интоксикации. Указанные характеристики представляют большой интерес для многих практических приложений: диспетчерские службы, кадровые агентства, медицина, психология и т.д. Достигнутые к настоящему времени успехи в этом направлении исследований являются несомненными, однако, окончательное решение далеко от завершения.

Цель данной статьи — изложение основных результатов исследований по автоматическому распознаванию эмоционального состояния диктора, проведён-

ных в [1]. При выполнении таких исследований возникают две основные трудности. Во-первых, отсутствует чёткое определение эмоции. А во-вторых, отсутствует однозначный ответ на вопрос о соотнесении акустических особенностей речи диктора с его эмоциональным состоянием. Всё это приводит к различиям в формах классификации эмоций и произвольной расстановке акцентов разными группами исследователей [2].

В современных системах определения эмоционального состояния диктора можно выделить следующие основные этапы обработки [3, 4]:

— вычисление информативных параметров речевого сигнала. Сюда обычно включают оценки мощности, частоты основного тона, формантных частот, спектральных и кепстральных коэффициентов и др.;

— вычисление функционалов от информативных параметров, таких как перцентили, экстремумы и их отношения, моменты высших порядков, коэффициенты регрессии и т.п.;

— классификация объектов. Наибольшее распространение в последнее время получили классификаторы на основе смеси нормальных распределений и метода опорных векторов [5].

Многие из статистических функционалов от параметров речевого сигнала, вычисляемых при формировании пространства информативных признаков, введены из априорных соображений о виде функции распределения используемого параметра. Такой подход не всегда может быть признан удовлетворительным. В данной работе предложено использование статистического критерия, отражающего сходство видов распределений исследуемой характеристики при решении задачи классификации эмоциональных состояний.

Описание корпуса эмоциональной речи

Обучение и тестирование алгоритма проводились на основе собранного корпуса эмоциональной речи на русском языке, который содержал записи диалогов и монологов (отдельных фраз и связанных текстов) со следующими характеристиками:

— диалоги:

- перечень эмоций: нейтральное состояние, грусть, страх, радость, гнев;
- количество дикторов: 31 женщина и 30 мужчин;
- общая длительность записей: примерно 1,5 часа;

— фразы:

- одни и те же фразы записаны в разном эмоциональном состоянии;
- перечень эмоций: нейтральное состояние, грусть, страх, радость, гнев;
- количество дикторов: 29 женщин и 30 мужчин;
- общая длительность записей: 4,0 часа;

— тексты:

- содержат плавный переход от нейтрального состояния к эмоциональному;
- перечень эмоций: нейтральное состояние, подавленность, грусть, тревога, страх, благодарность, радость, раздражение, гнев;
- количество дикторов: 31 женщина и 29 мужчин;
- общая длительность записей: 7 часов 45 мин.

В качестве дикторов привлекались профессиональные актёры и студенты театральных вузов. Родной язык дикторов — русский.

Запись эмоциональной речи осуществлялась в цифровом виде в звуковые файлы WAV-формата с частотой дискретизации сигнала: 16 кГц, разрядностью квантования 16 бит, ти-

пом кодирования — ИКМ и в режиме — моно. Среднее отношение сигнал/шум в записях, входящих в корпус, составило около 37 дБ.

Средние спектры мощности речи в различных психоэмоциональных состояниях для собранного корпуса и фонового шума показаны на рисунке 1.

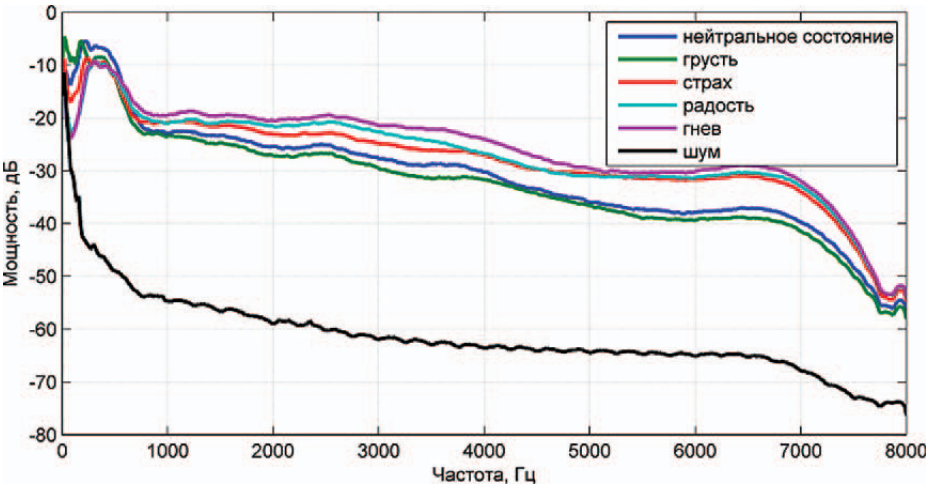

Рис. 1. Средние спектры речи в различных психоэмоциональных состояниях для собранного корпуса и фонового шума

Собранный речевой материал оценивался 12 экспертами. Каждый из них прослушивал полученные записи речи и определял, какую из эмоций выражает говорящий. В качестве оценки качества собранного речевого материала (меры согласованности оценок экспертов) использовался коэффициент Флейса:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e},$$

где $\overline{P}$ — средняя согласованность оценок экспертов, $\overline{P}_e$ — вероятность случайной согласованности оценок экспертов. Указанные величины определяются согласно следующим формулам:

$$\overline{P} = \frac{1}{N}\left[\sum_{i=1}^{N} P_i\right], \text{ где } P_i = \frac{1}{M(M-1)}\left[\sum_{j=1}^{K} m_{ij}(m_{ij}-1)\right];$$

$$\overline{P}_e = \sum_{j=1}^{K}\left[\frac{1}{NM}\sum_{i=1}^{N} m_{ij}\right]^2,$$

где N — общее количество классифицируемых записей речи конкретного диктора с имитацией заданной эмоции; P_i — степень согласованности оценок экспертов для каждой записи; M — количество экспертов; K — общее количество эмоций (равно 9), m_{ij} — количество экспертов, классифицировавших i-ю запись как относящуюся к j-й эмоции. Значение коэффициента P_i определяет следующие степени согласованности оценок экспертов, приведённые в табл. 1.

Записи речи, у которых степень согласованности оценок экспертов P_i менее 0,21, в итоговый корпус эмоциональной речи не включались. В результате коэффициент Флейса согласованности между оценками экспертов по всему итоговому корпусу получился равным **0,69** для фраз и **0,57** для текстов.

Таблица 1

P_i	Степень согласованности
0,01–0,20	Слабая
0,21–0,40	Посредственная
0,41–0,60	Средняя, умеренная
0,61–0,80	Значительная
0,81–1	Почти полная

Методика исследований

Информативные признаки

Исследования в области психолингвистики предоставили сведения о множестве акустических, просодических и лингвистических характеристик речи, способных служить информативными признаками при распознавании эмоционального состояния и проявляющихся на уровне голосовых сегментов, слогов и слов. При этом всё множество разбирают на базовые параметры и вычисленные для них функционалы. К таким базовым параметрам в настоящем исследовании были отнесены следующие: кратковременная мощность речевого сигнала, оценка частоты основного тона [7], джиттер (модуляция частоты основного тона), шиммер (модуляция амплитуды сигнала), коэффициенты линейных спектральных пар; кепстральные коэффициенты, вычисленные на основе коэффициентов линейного предсказания, мел-кепстральные коэффициенты, коэффициенты вещественного кепстра, фонетические функции на основе вычисления лог-спектрального расстояния, расстояния Итакуры-Саито и COSH-расстояния [8], оценки асимметрии и эксцесса распределения ошибки линейного предсказания сигнала [9], отношения мощностей спектра в формантных полосах, энергетический оператор Тигера в формантных полосах частот и критических полосах слуха [10]. К ряду базовых параметров добавлялись их первая и вторая производные по времени, применялись энергетический оператор Тигера и вычитание медианного значения.

Предварительные эксперименты показали, что большая эффективность распознавания эмоций достигается при использовании только вокализованных участков речи. Поэтому в дальнейшем использовались базовые параметры, вычисленные только на вокализованных участках речевого сигнала.

Расстояние между классами данных

Вычисление статистических критериев (расстояния) между функциями распределения информативных параметров позволяет непосредственно проводить классификацию данных (распознавание эмоций).

Наиболее естественным решением является вычисление расстояния между двумя многомерными плотностями распределения. Однако для построения многомерных плотностей распределения требуется большое количество обучающих данных. В противном случае классификатор может оказаться неустойчивым. Поэтому вместо многомерных распределений было решено использовать одномерные распределения базовых параметров речевого сигнала.

Выбор критерия при вычислении расстояния между двумя функциями распределения играет такую же важную роль, как и правильный подбор базовых параметров сигнала.

Для построения пространства расстояний желательно для всех параметров использовать единую функцию расстояния, не требующую априорных предположений о видах распределения сравниваемых величин.

Широко распространённым вариантом является применение критерия Колмогорова-Смирнова [11]:

$$d_{KS}(n,m) = \sup_x \left| F_n(x) - F_m(x) \right|,$$

где $F_n(x)$ и $F_m(x)$ — функции распределения параметра *x* в *n*-й и *m*-й записях соответственно. Неудобство использования этого вида расстояния состоит в том, что при сравнении двух практически не перекрывающихся распределений расстояние между ними будет равняться 1, вне зависимости от того, насколько удалены друга от друга эти распределения.

Поэтому в качестве функции расстояния было решено использовать критерий Джини [11]:

$$d(n,m) = \int \left| F_n(x) - F_m(x) \right| dx .$$

Таким образом, для каждого исследуемого параметра речевого сигнала необходимо вычислить критерий Джини между распределением этого параметра для анализируемого сигнала и распределением, априорно описывающим каждый класс эмоциональной речи. Эти расстояния целесообразно использовать как пространство наблюдений для обучения и тестирования классификатора.

При этом для описания функций распределения класса интуитивно подходящим представляется использование функции распределения всех данных этого класса. Однако при таком способе слабо учитывается поведение каждой эмпирической функции распределения, входящей в этот класс, и таким образом теряется информация о совокупности выборочных функций распределения класса как семейства соответствующих кривых.

Для преодоления этого недостатка функцию распределения класса целесообразно определить как медианное значение всех выборочных функций распределения, входящих в класс:

$$\widetilde{F}(x) = \operatorname*{median}_i \left(F_i(x) \right).$$

При этом определение медианной функции распределения целесообразно выполнять по равномерно расположенным квантилям всех наблюдений класса.

На рисунке 2 приведён пример различных способов формирования функции распределения класса: как функции распределения всех наблюдений (Class CDF) и как медианной функции распределения (Median CDF). На левом рисунке хорошо заметно наличие двух областей группирования функций распределения, связанных с различием распределений частот основного тона мужских и женских голосов. При этом следует отметить, что Median CDF, в отличие от Class CDF, в значительной степени сохраняет форму кривых выборочных функций распределения класса.

Окончательно формирование пространства наблюдений для классификации записей собранного корпуса эмоциональной речи формировалось из значений критерия Джини для каждого базового параметра и каждой медианной функции распределения классов. Т.е. размерность пространства наблюдений, подаваемого на классификатор, является произведением количества используемых базовых параметров речевого сигнала на количество анализируемый эмоций.

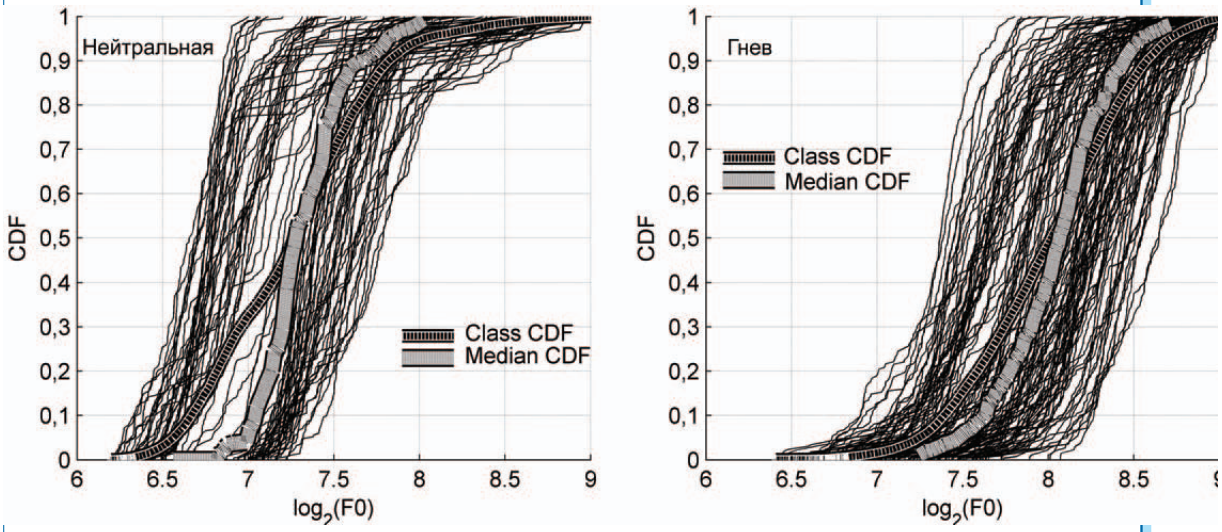

Рис. 2. Различные способы формирования функции распределения класса

Алгоритм классификации

Классификация объектов в сформированном пространстве признаков — завершающая операция большинства автоматических систем анализа речи. Для выполнения классификации был использован метод опорных векторов (SVM метод), реализованный в библиотеке libSVM [12]. В данной библиотеке мультиклассовый SVM-классификатор строится как набор классификаторов «каждый-с-каждым».

Это позволяет на этапе выбора информативных параметров сразу определить лучший набор именно для мультиклассовой классификации. Для построения разделяющей гиперповерхности в работе использовалось RBF-ядро [12], как наиболее универсальное и не требующее априорных предположений о характере распределения наблюдений. Эффективность распознавания в процессе определения оптимального набора информативных признаков и подбора параметров модели оценивалась при помощи метода K-кратной кросс-проверки реализованного в составе пакета libSVM.

Результаты экспериментов

Выбор количества информативных параметров

Для выбора множества информативных признаков использовалась стратегия линейного последовательного поиска, нашедшая широкое применение при решении задач распознавания эмоциональных состояний по голосу [3]. Суть её заключается в том, что на каждой итерации к множеству добавляются признаки, обеспечивающие наибольший прирост эффективности классификации. Чтобы увеличить гибкость алгоритма, на каждой итерации после добавления некоего подмножества из m признаков, обеспечивающего максимальный прирост эффективности классификации, производится удаление подмножества из n признаков. Нами использовались значения m = 5 и n = 2.

Зависимость оценки средней вероятности правильного автоматического распознавания пяти анализируемых эмоций от размерности множества информативных признаков показана на рисунке 3.

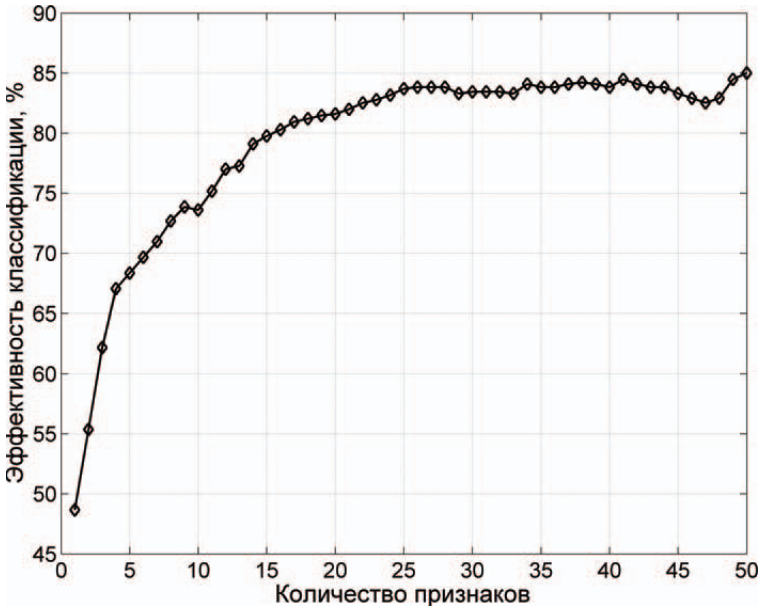

Рис. 3. Оценка средней вероятности правильного распознавания эмоций в зависимости от числа информативных признаков

Из рисунка видно, что на каждой итерации эффективность распознавания эмоций возрастала до тех пор, пока не достигла некоторого максимума, соответствующего примерно 25 информативным признакам. Дальнейшее повышение эффективности распознавания с ростом количества признаков происходило очень медленно.

Таким образом, для обучения классификатора использовался набор из следующих 25 информативных признаков:

— частота основного тона, её первая и вторая производные;

— первый, четвёртый и девятый коэффициенты линейных спектральных пар;

— производная первого и второго коэффициентов линейных спектральных пар;

— энергетический оператор Тигера от первого коэффициента линейных спектральных пар;

— оценки асимметрии и эксцесса распределения ошибки линейного предсказания сигнала;

— второй, третий и шестой коэффициенты вещественного кепстра (для значений частоты 125, 250 и 625 мс соответственно);

— первые производные третьего и четвёртого коэффициентов вещественного кепстра;

— первый, четвёртый, девятый, десятый и тринадцатый мел-кепстральные коэффициенты;

— первая производная первого и двенадцатого мел-кепстральных коэффициентов;

— вторая производная тринадцатого мел-кепстрального коэффициента;

— отношение мощности первой форманты (в полосе 150–850 Гц) к мощности четвёртой форманты (в полосе 2500–4000 Гц).

Оценка указанных параметров речи проводилась в полосе частот до 4000 Гц. Анализ сигналов выполнялся кадрами по 30 мс и с шагом 10 мс. Вычисления коэффициентов предсказания выполнялись для модели фильтра 12-го порядка [13]. На основе коэффициентов предсказания вычислялись 12 коэффициентов линейных спектральных пар в диапазоне от 0 до ϖ радиан и 16 кепстральных коэффициентов. Для ошибки линейного предсказания сигнала вычислялись оценки асимметрии и эксцесса. 13 мел-кепстральных коэффициентов вычислялись по 23 значениям спектра, равномерно распределённым по шкале мел-частот.

Результаты распознавания

Экспериментальная оценка эффективности автоматического распознавания классификации проводилась при помощи K-кратной кросс-проверки с K=10 [14]. В табл. 2 приведены экспериментальные оценки вероятностей правильного автоматического распознавания эмоций, полученные для 5 базовых эмоций: нейтральное состояние, грусть, страх, радость и гнев, а также соответствующие ошибки распознавания.

Из полученных результатов следует, что для собранного корпуса эмоциональной речи изложенный выше подход позволил обеспечить правильное автоматическое распознавание нейтрального состояния говорящего и 4-х указанных эмоций со средней вероятностью 84%.

Таблица 2

Результаты автоматического распознавания базовых эмоций для русского языка

Фактический класс	**Предсказанный класс**				
	Нейтральное состояние	Грусть	Страх	Радость	Гнев
нейтральное состояние	**97,2**	2,8	0	0	0
грусть	13,4	**86,6**	0	0	0
страх	2,2	6,6	**77,3**	8,2	5,7
радость	0	0	6,3	**73,5**	20,2
гнев	0	0	1,9	11,4	**86,7**

Заключение

Предложен и программно реализован алгоритм автоматического распознавания ряда эмоций человека по его речи на русском языке, осуществляющий одновременный анализ 25 параметров. Выбрана аналитическая мера, определяющая расстояние между векторами наблюдений для каждого вида эмоций говорящего.

Создан обучающий и тестовый корпус данных, содержащий записи речи профессиональных актёров и студентов театральных вузов (общей длительностью 13 часов) с нейтральным эмоциональным состоянием и проявлениями эмоций грусти, страха, радости и гнева.

С использованием этого корпуса получены оценки качества работы реализованного алгоритма. Экспериментально показано, что при длительности речевого сообщения не менее 5 секунд он обеспечивает правильное распознавание нейтрального состояния с вероятностью 0,97 и указанных эмоций — с вероятностью от 0,73 до 0,86, что в целом сравнимо с аналогичными результатами при распознавании эмоций человеком.

В качестве дальнейшей работы представляется целесообразным получение оценки эффективности реализованного алгоритма на корпусах речевых данных, содержащих записи речи с естественными проявлениями различных эмоций, поскольку в настоящей работе эти эмоции лишь имитировались.

Список литературы

1. *Исследование методов автоматического распознавания психоэмоционального состояния человека по его речи: Отчёт о НИР («Ш-2011–08–5.3»). ООО «ЦРТ», г. Санкт-Петербург, 2012.*

2. *The science of emotion: research and tradition in the psychology of emotions / Cornelius R. R // Prentice-Hall, Upper Saddle River: NJ, 1996.*

3. *Schuller B., Batliner A., Steidl S., Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge // Speech Communication. 2011. Vol. 53, № 9–10, P. 1062–1087.*

4. *Eyben F., Wöllmer M., Schuller B. OpenEAR-Introducing the Munich open-source emotion and affect recognition toolkit // Proc. IEEE ACII, 2009.*

5. *Bone D., Black M., Ming Li, Metallinou A., Sungbok Lee, Narayanan S.S. Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors // Proc. Interspeech, Florence, Italy, 2011. Pp. 3217–3220.*

6. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B. A Database of German Emotional Speech // Proc. Interspeech, Lisbon, 2005. P. 1517–1520.*

7. *Talkin D. A robust algorithm for pitch tracking (RAPT) // Speech coding and synthesis, Eds.: Elsevier Science, 1995. P. 495–518.*

8. *Rabiner L. R. Biing-Hwang Juang. Fundamentals of speech recognition: Prentice Hall. P. 1993–507.*

9. *Nemer E., Goubran R., Mahmoud S. Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain // IEEE Transactions on Speech and Audio Processing, 2001. Vol. 9, №. 3, P. 217–231.*

10. *Rahurkar M., Hansen J. H. L., Meyerhoff J., Saviolakis G., Koenig M. Frequency Band Analysis for Stress Detection Using a Teager Energy Operator Based Feature // Proc. Inter. Conf. on Spoken Language Processing ICSLP-2002, Denver, CO USA, 2002, Vol. 3, P. 2021–2024.*

11. *Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006.*

12. *Chang C.-C., Lin C.-J. LIBSVM: a library for support vector machines // ACM Transactions on Intelligent Systems and Technology, 2011, V. 2, № 27. P. 1–27.*

13. *Маркел, Дж.Д., Грэй А.Х. Линейное предсказание речи: Пер. с англ. // Под ред. Ю.Н. Прохорова и В.С. Звездина. М.: Связь, 1980.*

14. *Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection // Proc. of the Fourteenth International Joint Conference on Artificial Intelligence, 1995, Vol. 2. P. 1137–1143.*

Сведения об авторах:

Хитров Михаил Васильевич —
кандидат технических наук, генеральный директор и основатель Общества с ограниченной ответственностью «**Центр речевых технологий**», заведующий кафедрой **речевых информационных систем** (Национальный исследовательский университет информационных технологий, механики и оптики (**НИУ ИТМО**). Вице-президент консорциума «Российские речевые технологии», член ISCA, IEEE.
E-mail: khitrov@speechpro.com

Давыдов Андрей Геннадьевич —
кандидат технических наук, старший научный сотрудник Академии управления при Президенте Республики Беларусь. Основные научные интересы связаны с областью цифровой обработки сигналов, анализом, сжатием и распознаванием речи. Автор более 20 научных публикаций и 3 патентов.

Ткаченя Андрей Владимирович —
младший научный сотрудник ООО «Речевые технологии», г. Минск. Область научных интересов — системы анализа и индексирования аудиосигналов, скрытые Марковские модели в задачах распознавания речи.

Киселёв Виталий Владимирович —
директор ООО «Речевые технологии», г. Минск. С 1999 г. профессионально занимается системами синтеза и распознавания речи, диалоговыми речевыми системами. Автор более 25 научных публикаций в области речевых технологий. Основные научные интересы связаны с системами обработкой и анализом текста и речи, системами синтеза, распознавания речи, поиска ключевых слов.

Ромашкин Юрий Николаевич —
кандидат технических наук, окончил Московский инженерно-физический институт, факультет «Автоматика и электроника». Область научных интересов: цифровая обработка речевых сигналов, фильтрация речи на фоне помех, автоматическое распознавание речи и языка, идентификация говорящего по голосу, низкоскоростное кодирование речи, оценка качества трактов речевой связи.

E-mail: romayn@yandex.ru