УДК 621.391:004.934.2
doi:10.15217/issn1684-8853.2017.5.30

# АНАЛИТИЧЕСКИЙ ОБЗОР КОМПЬЮТЕРНЫХ ПАРАЛИНГВИСТИЧЕСКИХ СИСТЕМ ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ ЛЖИ В РЕЧИ ЧЕЛОВЕКА

***А. Н. Величко***[a, б], *магистрант*
***В. Ю. Будков***[a], *канд. техн. наук*
***А. А. Карпов***[a], *доктор техн. наук, доцент*
[a]*Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, РФ*
[б]*Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, РФ*

***Постановка проблемы:*** *компьютерная паралингвистика анализирует невербальные аспекты человеческой коммуникации и речи, такие как естественные эмоции, интонации, особенности произношения, параметры голоса диктора, истинность речевых сообщений и т. д. Задача автоматического выявления истинности/ложности сообщений является актуальной в различных приложениях, многие современные исследования посвящены разработке математического и программного обеспечения для автоматизированных систем распознавания лжи в речи человека.* ***Цель:*** *анализ и представление достижений и разработок в области компьютерной паралингвистики, в частности, в автоматическом распознавании лжи в речи человека для определения недостатков существующих методов и путей их преодоления при создании новой автоматической системы.* ***Результаты:*** *анализ широкого спектра современной научно-технической литературы, описывающей результаты мировых научных исследований по данной тематике за последние десять лет, включая международные соревнования Computational Paralinguistic Challenge, показал, что применяются во многом схожие методы распознавания, однако алгоритмы обработки сигналов имеют различия, которые влияют на точность распознавания ложности/истинности речевых высказываний. Представлена обобщенная схема системы распознавания, ее основные составляющие, а также классификация наиболее эффективных методов, использующихся при разработке автоматических систем паралингвистического анализа естественной речи. На данный момент в распознавании лжи в речи человека существует масса нерешенных проблем технического и естественного характера, включая учет индивидуальных особенностей диктора (его пол, возраст, эмоциональную стабильность, национальные особенности и т. д.), преодоление которых позволит значительно улучшить функциональность системы.*

***Ключевые слова*** *— компьютерная паралингвистика, речевые технологии, распознавание лжи в речи человека, машинное обучение.*

## Введение

Паралингвистика — область науки, которая изучает невербальные аспекты человеческой коммуникации и речи: естественные эмоции, интонации, акценты, психофизиологические состояния, особенности произношения, параметры голоса диктора, ложность или истинность речевых сообщений и т. д. В основном современная паралингвистика рассматривает то, как произносится речь, нежели то, что произносится [1].

Хорошо известен факт, что наше физиологическое состояние очень тесно связано с эмоциональными переживаниями. Идея детекции лжи по речевому сигналу основывается на гипотезе о том, что ложь вызывает у человека состояние стресса, что и отражается на изменении параметров речи. Эффект Липпольда [2] заключается в том, что все мышцы человека, в том числе и голосовые связки, подвержены микроколебаниям с частотой 8–12 Гц, при этом в спокойном состоянии частота этих колебаний не превышает 10 Гц, а в стрессовом возрастает до 12 Гц.

С развитием технологий, позволяющих распознавать речь человека, многие организации проявляют интерес к данной области, поскольку в современном мире достаточно остро стоит проблема распознавания лжи в речи человека. Ложь — это преднамеренный акт введения собеседника в заблуждение посредством передачи неверной или вводящей в заблуждение информации [3]. Ложная информация бывает преднамеренной (дезинформация) и непреднамеренной (заблуждение). Помимо отличий между самоориентированной ложью и ложью, ориентированной на других, часто приводится различие между явной ложью (полная ложь, диаметрально противоположная истине), преувеличением (сообщаемая информация или факты превосходят истинные данные) и тонкой ложью (сообщение практически истинно, но составлено для заблуждения; уклонение от ответа или умышленное опущение деталей).

Тема распознавания ложных речевых сообщений становится особенно актуальной, поскольку на данный момент большинство исследований на тему лжи опираются на визуальном ее проявлении, т. е. на мимике, жестах, биометрических параметрах, что можно распознать при исследованиях с использованием полиграфа [4, 5].

Несмотря на популярность использования полиграфа, этот метод не является оптимальным, поскольку предъявляются особые условия для работы с аппаратом — и к месту исследований (комфортный температурный режим, оптимальная влажность, шумоизоляция и пр.), и к испытуемому (в первую очередь, наличие добровольного согласия на проведение испытаний, отсутствие соматических заболеваний, психических расстройств и пр.). Именно поэтому возникла заинтересованность в методах, не подразумевающих физический контакт с испытуемым, а именно в бесконтактных методах, исследующих речевую активность и невербальные сигналы. Однако стоит отметить, что данная задача является комплексной ввиду многих факторов, влияющих на анализ звукового сигнала: неоднозначность языка, индивидуальные особенности диктора (дефекты слуха, речи и пр.), наличие шумов при записи.

## Методы автоматического анализа паралингвистических явлений в речи

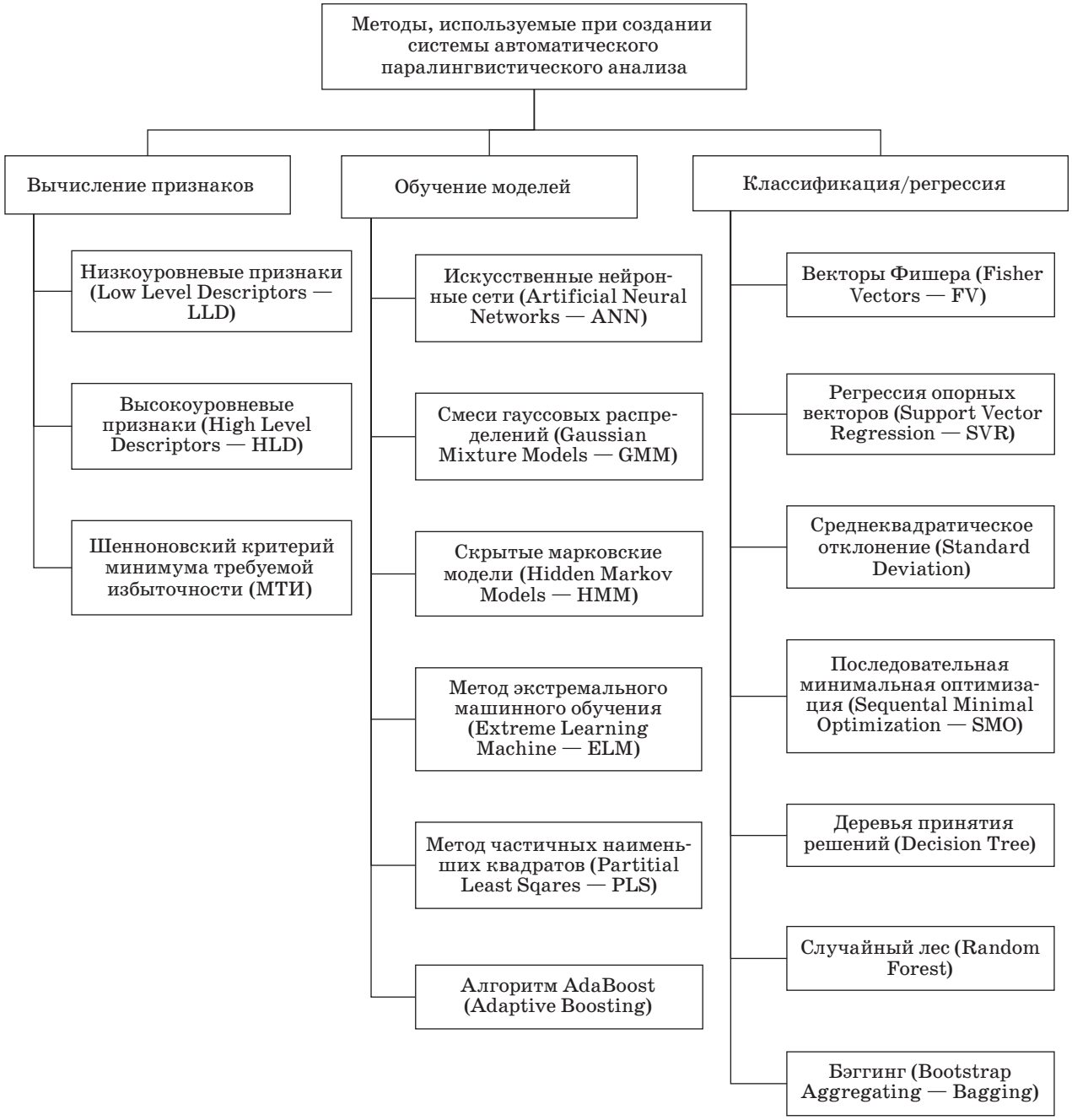
В современных системах паралингвистического анализа речи используются пространства признаков огромного размера (низкоуровневые описатели, Low Level Descriptors — LLD) для интегрального описания фраз, а не отдельных слогов и фонем (супрасегментные признаки). Эти же LLD-признаки используются для описания речевых сигналов в компьютерной паралингвистике [1, 3, 6]. Обычно наборы признаков включают в себя частоту основного тона (ЧОТ), форманты (резонансные частоты голосового тракта), мел-частотные кепстральные коэффициенты (Mel-Frequency Cepstral Coefficients — MFCC), модулированный спектр сигнала, коэффициенты перцептивного линейного предсказания (Relative Spectral Transform — Perceptual Linear Prediction — RASTA-PLP), энергетические признаки сигнала и их вариативности (так называемые джиттер и шиммер) и т. д. MFCC- и RASTA-PLP-признаки известны в автоматическом распознавании речи довольно давно и были привнесены в распознавание паралингвистических явлений речи из этой задачи, недавно к ним добавились также частотные признаки речи (Line Spectral Frequency — LSF) [1]. Однако многие исследователи экспериментируют с наборами признаков, включая в них и другие признаки. Такие наборы признаков представлены, например, в работах [7, 8].

Наиболее распространенными методами моделирования и классификации паралингвистических явлений на сегодняшний день являются: искусственные нейронные сети, векторы Фишера (Fisher Vectors — FV), смеси гауссовых распределений (Gaussian Mixture Models — GMM), регрессия опорных векторов, скрытые марковские модели (Hidden Markov Models — HMM), модель экстремального машинного обучения (Extreme Learning Machines — ELM), метод частичных наименьших квадратов, последовательная минимальная оптимизация (Sequental Minimal Optimization — SMO), случайный лес (Random Forest), бэггинг (Bagging, Bootstrap Aggregating), деревья принятия решений, среднеквадратичное отклонение, шенноновский критерий минимума требуемой избыточности (МТИ) [9, 10]. На рис. 1 приведен вариант классификации указанных методов.

При разработке математического и программного обеспечения паралингвистических систем многие исследователи рассматривают вопрос об использовании свободно доступного программного обеспечения для проведения экспериментов. При необходимости определить эмоциональное состояние личности может быть полезным такой программный продукт, как LIWC (Linguistic Inquiry and Word Count, https://liwc.wpengine.com), который является условно бесплатным программным обеспечением для анализа текстов, вычисления частотности использования слов человеком, определения эмоциональной нагрузки текста. Для вычисления признаков можно применять инструментарий openSMILE (http://audeering.com/technology/opensmile/), использующийся во многих работах для извлечения признаков из аудиозаписей.

Набор акустических низкоуровневых признаков (LLD) в openSMILE состоит из 6373 супрасегментных признаков, включая 65 базовых низкоуровневых признаков, а также их варианты (рис. 2). Низкоуровневые признаки включают в себя множество характеристик: спектральные, кепстральные, энергозависимые и вокализованные. Эти акустические признаки считаются наиболее полными для паралингвистических исследований.

В качестве программных средств, реализующих алгоритмы извлечения и анализа данных, можно применять программный инструментарий WEKA (Waikato Environment for Knowledge Analysis, www.cs.waikato.ac.nz/ml/weka/), представляющий собой набор средств визуализации и алгоритмов для анализа данных и решения задач прогнозирования. Для фонетического анализа речи можно использовать инструментарий Praat (www.fon.hum.uva.nl/praat/). В качестве программных продуктов для анализа данных можно применять такие продукты, как KNIME и RapidMiner. KNIME (Konstanz Information Miner, www.knime.org) представляет собой систему построения алгоритмов для анализа, преобразования и визуализации данных. Может интегрироваться с другими проектами, например с WEKA.

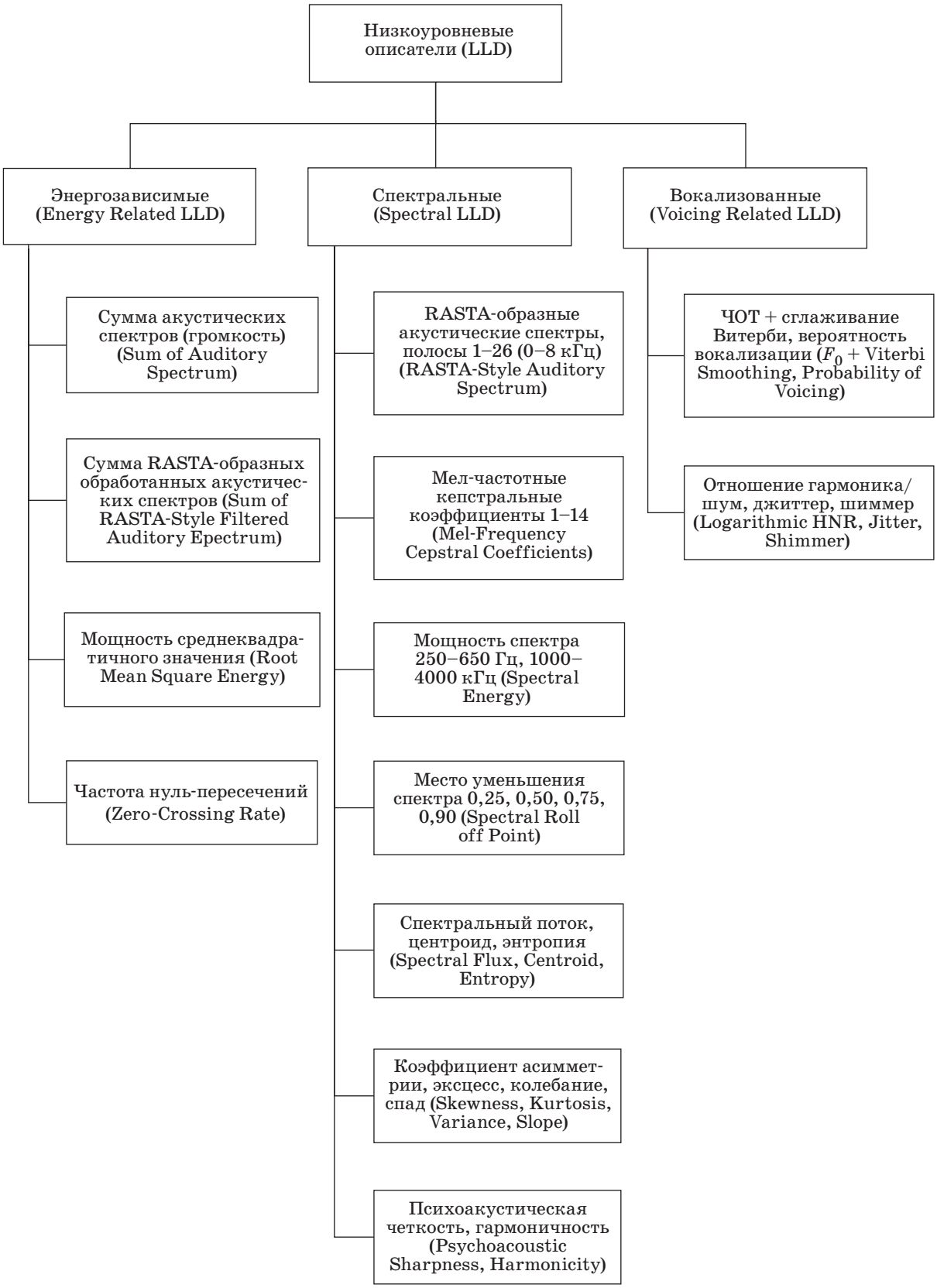



■ ***Рис. 1.*** Классификация основных методов паралингвистического анализа речи
■ ***Fig. 1.*** Classification of major methods for paralinguistic speech analysis

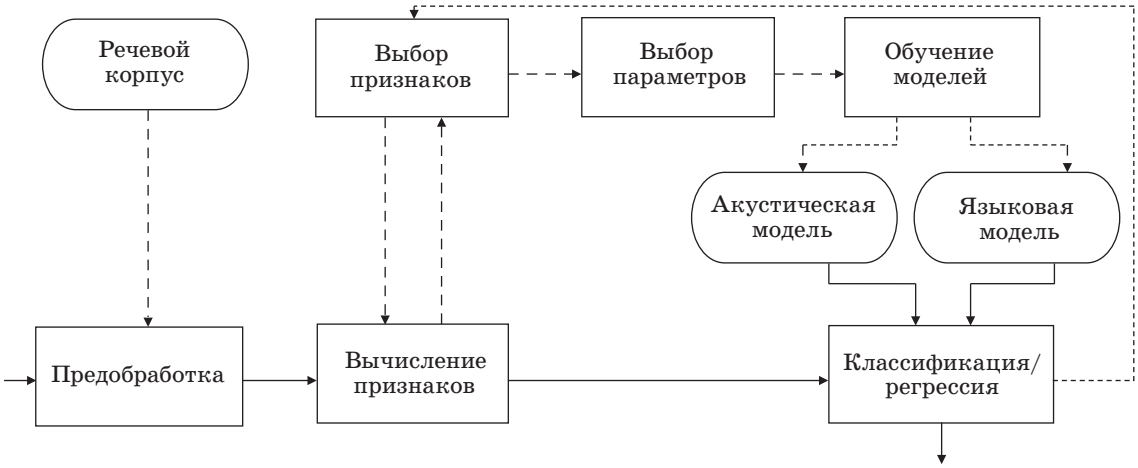
RapidMiner (https://rapidminer.com) — открытый программный продукт для проведения экспериментов в таких сферах, как машинное обучение и интеллектуальный анализ данных. Он может быть использован для интеллектуального анализа данных, текста, мультимедиа, потоков данных. Программный продукт интегрирует в себя операторы WEKA, имеет встроенный язык сценариев для выполнения массивных серий экспериментов.

Вышеописанные методы являются компонентами классической системы паралингвистического анализа речи, общая схема и основные этапы работы которой представлены на рис. 3 [1].

Для того чтобы построить модель паралингвистического анализа речи, необходимо собрать или подготовить речевую базу данных, на которой будет проводиться обучение моделей, тестирование (отладка) системы и классификация паралингвистических явлений. При сборе материа-




■ ***Рис. 2.*** Низкоуровневые признаки в инструментарии openSMILE
■ ***Fig. 2.*** Low level descriptors in openSMILE toolkit




■ ***Рис. 3.*** Обобщенная схема системы паралингвистического анализа речи
■ ***Fig. 3.*** General scheme of a paralinguistic speech analysis system

ла для обработки речи, в том числе и содержащей ложь, существуют проблемы, связанные с анализом персональных данных диктора (пол, возраст, образование, эмоциональная стабильность/нестабильность, культурные и национальные отличия). Базы данных собираются, в основном, в изолированных условиях (в лаборатории), а не в естественных, что значительно влияет на качество речи (преувеличение, неточное отображение эмоций). При этом желательно, чтобы количество материала было достаточным для успешного проведения всех трех стадий работы системы (обучение, тестирование, классификация).

## Речевые базы данных с ложными речевыми сообщениями

Известны несколько общедоступных корпусов речи, содержащих как ложные, так и правдивые речевые сообщения дикторов. К примеру, в работе [3] использовался речевой корпус DSD (Deceptive Speech Database), разработанный в Университете Аризоны (США). Он состоит из аудиозаписей, включающих в себя 162 минуты речи 72 дикторов. В записи участвовали студенты университета, которые были поделены на две группы. Участники первой группы играли роль лжецов, которые «украли» ответы на экзамен из компьютера на кафедре. Вторая группа играла роль честных учеников, которые вернули листовку в тот же кабинет. Следующая фаза заключалась в том, что были проведены интервью с каждым участником. Участники, которые украли ключ, должны были лгать, что они этого не делали, в течение всего интервью, другая группа должна была говорить правду о своих действиях. Интервью состояло из подготовленного набора открытых вопросов, подразумевающих короткие ответы (десять «фоновых» вопросов для основы, специальные вопросы о краже).

В работе [11] представлен корпус эмоциональной речи GEMEP (Geneva Multimodal Emotion Portrayals), который включает коллекцию аудио- и видеозаписей, содержащих информацию от 10 франкоговорящих актеров (5 женщин, 5 мужчин), которые изображают 18 эмоциональных состояний (восторг, изумление, злость, чуткость/отзывчивость, отвращение, отчаяние, гордость, стыд, беспокойство, интерес, радость, презрение, панический страх, удовольствие, облегчение, удивление, грусть), применяя различные вербальные средства и различные степени выражения (сильную или слабую).

В работе [12] использовался меньший по размерам речевой корпус — CSC (Columbia-SRI-Colorado), разработанный в Университете Колумбии (США) и состоящий из 32 часов аудиозаписей интервью 32 носителей стандартного американского языка (16 мужчин, 16 женщин). Организаторами было проведено интервью с каждым участником, также участникам было предложено выполнить серию заданий. Им было сказано, что их результаты сравнят с характеристикой одного из ведущих бизнесменов Америки. После им выдали подтасованные результаты и попросили сыграть снова, чтобы достичь наиболее близких результатов с заданной характеристикой. В четырех из шести заданий участники обманули интервьюеров.

В работе [13] авторы использовали корпус, собранный в Университете Ноттингема (Великобритания) при участии 19 мужчин (студентов и преподавателей). У всех участников родной язык — английский, дикторы не имели каких-либо отклонений (слуховых, речевых и пр.). Пример опроса с оценкой был разработан для этого эксперимента. Участникам были выданы жетоны с тек-

стовой информацией о том, что они должны скрывать от интервьюера во время оценочного опроса. Интервью 1 с базовыми данными состояло из нейтральных и расслабляющих вопросов, составленных для того, чтобы иметь материал с истинными сообщениями. Следующие два интервью были составлены таким образом, чтобы задать как можно больше наводящих вопросов. Интервью 2 вызвало существенные затруднения при постановке обычных вопросов о социальной привлекательности и сокрытии информации. Интервью 3 было более провокационным из-за прямого опроса участников об их честности.

Авторы работы [14] считают, что поведение лжеца напрямую зависит от индивидуальных качеств, культурного уровня человека, содержания разговора и цены наказания в случае разоблачения. Именно поэтому данные, содержащие ложные сообщения, должны быть собраны в реальных условиях. Авторы придумали интересную игру (сценарий), в ходе которой был создан речевой корпус на мандаринском китайском языке. Есть две группы (А и Б). Каждый член группы А должен рассказать историю (реальную или блеф), а члены группы Б могут задать любые вопросы по этой истории. Истории членов группы А были разными, соответственно, задавались разные вопросы и получены разные ответы на них. Учитывая тот факт, что члены группы Б не знали, действительно ли автор пережил то, что рассказал в истории, они должны были решить, правда это или ложь, по ответам рассказчика. Если члены группы Б угадывали ответ, они выигрывали игру и получали награду, в противном случае выигрывала группа А. Если история — ложь, рассказчик должен сделать все возможное, чтобы скрыть это от остальных, чтобы выиграть игру. Члены группы Б могли задавать сколько угодно вопросов в надежде, что рассказчик начнет нервничать и путаться в событиях. Авторы отобрали каждую историю-блеф и фальшивые ответы как данные с ложными сообщениями. Затем была записана нейтральная речь людей, которым принадлежали эти истории, в нормальных условиях. Тематика разговоров могла включать представление себя, рассказ о хобби, жизни и прочее. Записи должны были быть длинными настолько, чтобы включить в них как можно больше слогов китайского мандаринского диалекта. В итоге авторы получили 50 записей участников, включающих речь 25 мужчин и 25 женщин 25–35 лет.

### Экспериментальные системы распознавания лжи в речи человека

В рамках международной конференции INTERSPEECH с 2009 г. проходят соревнования (де-факто чемпионат мира) по различным направлениям компьютерной паралингвистики Computational Paralinguistics Challenge (ComParE). В 2016 г. впервые на соревнованиях появились следующие темы: распознавание лжи в речи, распознавание степени искренности человека, а также идентификация родного языка диктора по его англоязычной речи [15]. Исследователи, участвующие в данном соревновании, могли использовать собственные алгоритмы машинного обучения и наборы признаков в дополнение к представленному организаторами соревнований стандартному набору признаков (6373 признака, вычисленных посредством openSMILE). Также участникам предоставлялись обучающие/отладочные аудиозаписи. Для конкурса распознавания лжи и степени искренности в речи в качестве базового критерия оценки результатов применялся количественный показатель UAR (Unweighted Average Recall — среднее значение полноты). Среднее значение полноты — это мера измерения, которая лучше других подходит для данных с несбалансированными классами, поскольку она основывается на средней чувствительности и специфичности (mean of sensitivity and specificity). Преимущество ее в том, что она взвешивает каждый класс независимо от количества субъектов, которые он содержит. Пусть **A** — матрица ошибок (contingency matrix), где $A_{ij}$ — это число субъектов класса $I$, который классифицируется как $j$, и пусть $K$ будет количеством классов, тогда

$$UAR = \frac{1}{k}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}}.$$

В качестве речевого корпуса для соревнования была предложена база данных ложной речи DSD. Организаторы провели собственные испытания, где использовали свободное программное обеспечение openSMILE и WEKA, с помощью которых были выделены признаки, вошедшие в базовый набор параметров, а также предоставлена базовая реализация алгоритмов обработки аудиосигналов. Организаторы предоставили результаты испытаний базовой системы, которая показала значения UAR на отладочном и тестовом наборах данных 61,9 и 68,3 % соответственно, что далее использовалось в качестве минимальной планки [15].

Для участия в соревнованиях было зарегистрировано более 20 команд из разных стран мира, которые представили свои компьютерные паралингвистические системы для выявления лжи в речи, а также описывающие их статьи. Авторы работы [3] использовали базу данных DSD для того, чтобы с помощью программного обеспечения для распознавания речи CMU Sphinx (Carnegie

Mellon University) определить лингвистические составляющие: просодические признаки и типы ответов. Из транскрипций речи были вычислены два типа просодических признаков: обычные меры оценки речи и аудиопризнаки, основанные на парадигме переменной разрешающей способности. В ходе экспериментов с применением вычисленных признаков был получен результат 74,9 %.

В другой работе соревнования [6] был предложен новый набор признаков, состоящий из акустико-просодических, лексических, синтаксических и фонотактических признаков. Также было проведено оценивание каждого признака на полезность его для данного задания. В работе были использованы корпуса CSC и DSD. На этапе разработки при использовании предложенного набора признаков и базового набора соревнований получены результаты 67,7 и 62,2 % соответственно.

Некоторые исследователи предполагают [9–11], что ложь в речи может быть определена с помощью эмоциональных признаков. Эмоциональная насыщенность речи проявляется в темпе, тембре и громкости речи. О безразличии к излагаемой информации свидетельствует вялая и неэмоциональная речь. Если темп речи высокий, то можно предположить, что человек взволнован, тема разговора его глубоко волнует, он как бы пытается сказать больше, чтобы убедить в своей правоте. Тембр голоса — это характер звука голоса, его окраска, через которые также выражается отношение к теме разговора. Таким образом можно подробнее узнать личность говорящего: лексические признаки говорят об образовании, социальном статусе и возрасте; грамматические признаки свидетельствуют о том, насколько грамотен и образован человек; синтаксические признаки могут подчеркивать как недостаточные навыки построения фраз, так и чрезмерное возбуждение; стилистические признаки отражают навыки использования речи при общении.

Авторы работы [11] предположили, что ложь можно распознать, используя такие признаки эмоциональности, как интенсивность/сила эмоции (arousal), валентность/тон эмоции (valence), эмоциональная регуляция. При этом обучение проходило на речевом корпусе GEMEP при помощи классификатора k-ближайших соседей (k-Nearest-Neighbour — kNN). Для основной системы авторы совместили методы kNN, SVM (Support Vector Machine — метод опорных векторов), SMO. Для получения окончательных результатов система была опробована на корпусе соревнований DSD. Были проведены исследования с использованием разработанного набора признаков, в результате которых комбинация наборов эмоциональных признаков и базовых LLD-признаков (ComParE-2013) показала значение UAR 68,9 %.

В работе [7] исследовался потенциал информативных признаков, основанных на автоматическом распознавании фонов в речи (фон — минимальная звуковая единица речи, рассматриваемая вне связи с функцией смыслоразличения). Транскрипции речи были использованы для обработки звуков (фонем, пауз тишины, заполненных пауз) и соответствующих им длительностей. Авторы предложили высокоуровневый набор признаков, в который входят четыре группы: гласные, фоны, псевдослоги и паузы. Из них отобрали 29 статических признаков и соответствующие им длительности, скорость речи и соотношения. Также выбрали подходящий, предложенный организаторами соревнования, набор акустических признаков и совместили эти два набора. Для работы был выбран корпус DSD, методы SVM с функцией линейного ядра (linear kernel function), SMO. Отдельно от набора признаков соревнований признаки, основанные на автоматическом распознавании фонов в речи, показали результат 58,6 %. Однако вместе с ним результаты UAR оказались 66,7 % на этапе разработки и 69,3 % — на этапе тестирования.

В работе [16] авторы предложили использовать FV для описания низкоуровневых признаков в высказываниях. Преимущество векторов Фишера в том, что этот метод требует намного меньшего количества составляющих в смеси гауссовых распределений, чем, например, модель «мешка слов» (Bag-of-Words — BoW) и не требует обучения на очень большом корпусе, как в случае универсальной фоновой модели (Universal Background Model — UBM). Кроме того, была использована каскадная нормализация и ELM. Наилучшие результаты UAR составили 75,2 и 66,6 % для отладочного и тестового наборов данных соответственно.

Авторы работы [17] предположили, что существует связь между распознаванием лжи и пониманием степени искренности в речи. Они продемонстрировали подход к решению этой задачи: объединение корпусов из заданий соревнования для определения лжи и искренности. Авторы сделали вывод, что метод, базирующийся на том, что ложь и искренность — «две стороны одной монеты», не показал ожидаемых результатов. Однако они смогли использовать взаимодействие между этими явлениями. Для экспериментов был выбран набор признаков соревнования и использован алгоритм SMO для обучения SVM. С помощью метода маркировки ошибочных данных авторы показали, что самообучающийся классификатор способен выбирать подходящие примеры из другого задания, которые дополняют уже существующие данные.

В рамках конкурса также были представлены работы по улучшению алгоритма для системы рас-

познавания лжи в речи. Авторы работы [18] предложили подход с использованием поверхностной нейронной сети (Shallow Neural Networks) в двух вариантах архитектуры: выборочная регрессия и ранжированная нейронная сеть (sampling-based regression and ranking neural networks) и анкерная регрессия и ранжированная нейронная сеть (anchor-based regression and ranking neural network). Также они описали способ, который одновременно минимизирует регрессию и потери в классификации. Был использован набор признаков, предоставленный организаторами конкурса. В результате экспериментов обе архитектуры нейронных сетей, предложенные исследователями, показали потенциал.

Известен также ряд работ по определению лжи в речи, выполненных до соревнований INTERSPEECH 2016 Computational Paralinguistic Challenge. Так, в работе [12] сделан фокус на психологической стороне лжи, а не на визуальном или биометрическом ее проявлении. Автор считает, что главной проблемой изучения лжи является то, что на речь диктора может влиять множество факторов. Дикторы могут испытывать различные эмоции в зависимости от причины, по которой они лгут. Возраст и культура человека также играют важную роль. Было описано исследование с использованием деревьев принятия решений для предсказания лжи. Были вычислены признаки, включающие слоги, слова, предложения, короткие предложения и «простые» предложения; эталоны слов и сложности предложений; индикаторы спецификации и выразительности; «неформальные» эталоны, основанные на ошибках, которые были автоматически распознаны. Лучшие результаты с деревом принятия решений были получены из 20 перекрестных проверок, запущенных на малом наборе данных, они показали точность 70 %. В ходе исследования был разработан корпус CSC, описанный выше.

В работе [13] авторы изучили свойства речи людей, которые лгали во время опроса. В эксперименте участвовало 19 человек. Собранные данные были проанализированы с использованием диапазона параметров речи, включая скорость речи (Speaking Rate — SR), время начала ответа (Response Onset Time — ROT), частоту и длительность пауз. Полученные результаты показали заметное ускорение темпа речи, уменьшение времени начала ответа, уменьшение длительности хезитаций во время произнесения лжи.

Авторы работы [14] описали применение дробного преобразования Фурье ( Fractional Fourier Transform — FrFT) для получения признаков, необходимых для определения ложности сообщения. Если психоэмоциональное состояние человека влияет только на частотные характеристики речевого сигнала, то можно использовать обычное оконное преобразование Фурье, но его будет недостаточно, если в сигнале изменяется только фаза. Поэтому вместо стандартных MFCC-признаков в данной работе были применены дробные мелкепстральные коэффициенты (Fractional Mel Cepstral Coefficient — FrCC). 25 мужчин и 25 женщин были участниками эксперимента, результаты которого показывают, что при выборе оптимального порядка FrFT для FrCC-признаков точность распознавания ложных речевых сообщений выше, чем при применении MFCC-признаков, при этом FrCC лучше кластеризуются. При использовании модели линейного дискриминантного анализа (Linear Discriminant Analysis — LDA) и FrCC-признаков средняя точность распознавания ложных речевых сообщений для мужчин и женщин составила 59,9 и 56,2 % соответственно. Для MFCC-признаков точность была меньше на 3,2 и 5,9 % соответственно. Использование HMM увеличило точность до 71 и 70,2 %. Результаты показали, что ложная информация действительно может быть выявлена из речевого сигнала с довольно высокой точностью.

Авторы работы [8] определяли ложь, используя как акустические, просодические и лексические признаки речи диктора, так и информацию о поле диктора, его этнической принадлежности и личностных факторах. Был собран корпус из записей диалогов 126 пар испытуемых, в сумме составляющий 93,8 часа речи. В ходе экспериментов были опробованы два метода: случайный лес и бэггинг. Авторы объединили методы случайного леса и метода J48 (имплементация алгоритма дерева решений C4.5 (C4.5 decision tree algorithm)). Было вычислено 14 акустико-просодических признаков для данного эксперимента: минимум, максимум, середина, медиана, среднеквадратичное отклонение значения ЧОТ, средний абсолютный спад, минимум интенсивности, максимум интенсивности, средняя интенсивность, среднеквадратичное отклонение интенсивности, дрожание, шумы, коэффициент гармоники и шума. Авторы использовали два метода нормализации. Первый заключается в нормализации признаков диктора с использованием среднего значения ЧОТ и среднеквадратичного отклонения признаков диктора в течение обеих частей сессии. Авторы назвали это сессией нормализации. Второй метод заключается в применении данных основной части эксперимента, в которой собраны 3–4-минутные фрагменты речи каждого участника (правдивые) до того, как они встретили партнера. Для того чтобы уловить отклонение диктора от произнесенной правды, авторы нормализировали речь дикторов во время произнесения лжи, используя признаки, выделенные из исходных данных. Авторы предположили, что эта нормализация, названная «нормализацией исходных

данных», будет полезной для распознавания лжи. Они использовали метод z-нормализации для обеих нормализаций. После проведения испытаний выяснилось, что метод случайного леса оказался более точным в определении лжи и показал точность распознавания 61,23 % на сыром наборе акустико-просодических признаков, 63,03 % при использовании сессионной нормализации и 32,79 % при использовании нормализации исходных данных, в то время как метод бэггинга при тех же испытаниях показал результаты 58, 65, 61, 19, 31, 01 % соответственно. Затем авторы совместили сессионную нормализацию, показавшую лучшие результаты на первом испытании, результаты теста по личностному пятифакторному опроснику NEO-FFI (Neuroticism-Extraversion-Openness Five-Factor Inventory), а также информацию о поле и родном языке дикторов. В ходе второго эксперимента результаты применения методов для распознавания лжи улучшились: точность распознавания составила 65,86 % при использовании метода случайного леса и 63,9 % при использовании метода бэггинга.

К изучению влияния эмоциональных признаков на ложь также обращались авторы работ [9, 10]. «Фонетический детектор лжи» является официально зарегистрированной программой для ЭВМ [10], которая предназначена для тестирования эмоционального состояния личности по голосу. Ее основной принцип действия — автоматическая оценка качества речи диктора на базовом, фонетическом, уровне по общесистемному шенноновскому критерию МТИ речевого сигнала. Авторы считают, что проблема современных систем, основывающихся на принципе последовательного членения голосового сигнала на короткие отрезки данных и их последующем сопоставлении с эталоном, в том, что они не учитывают человеческий фактор: диктор (в силу особенностей слуха или речи) может быть не в состоянии воспроизвести эталон, к тому же один и тот же диктор не всегда может воспроизвести эталон несколько раз одинаково. Для решения данной задачи было предложено записывать несколько эталонов.

В работе [19] были выявлены индивидуальные различия: некоторые люди повышают высоту голоса (ЧОТ), когда лгут, тогда как другие, наоборот, понижают; некоторые склонны к смеху во время произнесения лжи, другие же смеются, когда говорят правду. Также было определено, что помимо акустико-просодических признаков работу классификации улучшают дополнительные данные: пол, родной язык, персональные данные. Для исследования использовался корпус, часть которого была специально размечена для данной задачи. Корпус состоит из диалогов, длящихся 3–4 мин, где испытуемые отвечали на простые открытые вопросы. Из него были выделены акустико-просодические и лексические признаки, после чего проводилось обучение классификаторов определению пола, родного языка и личности говорящего. Также был проведен тест NEO-FFI, который определил открытость, доброжелательность, добросовестность, эмоциональность и экстраверсию участников. Для обработки сигнала был использован программный инструментарий Praat, а для выделения лексических признаков — LIWC. Исследователи использовали различные методы машинного обучения, и наиболее успешным оказался AdaBoost в сочетании с акустико-просодическими признаками и LIWC-признаками, был достигнут результат 61 % точности распознавания лжи.

В работе [20] представлена специализированная методика для выявления параметров речевого сигнала, отражающих истинность передаваемой информации. Использовались такие признаки, как: наличие вокализации звуков, ЧОТ, интенсивность основного тона, динамика изменения и девиация ЧОТ, динамика изменения интенсивности основного тона, отношение интенсивности гармоник к интенсивности основного тона. Для проведения исследования был разработан сценарий, состоящий из последовательности вопросов (нейтральные, контрольные, значимые). Для каждого ответа производится расчет признаков, после чего признаки значимых и контрольных ответов сравниваются и оцениваются следующим образом: присваивается значение 0 баллов, если различий в реакции нет; 1, если выявлены заметные различия; 2, если различия сильные; 3, когда различия ярко выражены. В случае если реакция на значимый вопрос была сильнее, чем на контрольный, балл принимает отрицательное значение, и наоборот, если реакция на контрольный вопрос сильнее, чем на значимый, то ставится положительная оценка. Результат общего теста авторы получали суммированием всех оценок. Полученные результаты показали, что вполне возможно использование данной системы для определения ложности в речи человека в режиме реального времени, в процессе межличностного общения.

Авторы патента РФ зарегистрировали способ определения искренности/неискренности по результатам трехкратной оценки эмоционально-психологических свойств и состояний человека по одному и тому же фрагменту видеозаписи методом психологического шкалирования [21]. В основе предложенного способа лежит идея о том, что при искреннем высказывании говорящий использует комплекс всех форм и элементов невербального поведения (мимика, жесты, интонации голоса — все работает согласованно). На первом этапе эксперты (не менее 10 человек, об-

ладающих эмоциональным слухом не менее 80 % по тесту В. П. Морозова, а также знающие основы выразительных движений человека) оценивают только по аудиосигналу, на втором — только по видеосигналу, на третьем — по обоим сигналам одновременно. Далее вычисляется коэффициент соответствия средних значений оценок экспертов при использовании вычисления ранговой корреляции по Спирмену. Оценку искренности/неискренности производят путем усреднения оценок всех психологических свойств и состояний говорящего.

## Заключение

Наличие многочисленных работ на тему определения лжи в речевых сообщениях свидетельствует о том, что на сегодняшний день тема является актуальной для различных приложений, например: предотвращение «телефонного терроризма»; биометрические исследования на полиграфе, проводимые правоохранительными органами и специальными службами; анализ поведения абонентов и операторов в диалоговых системах, установленных в контакт-центрах, в банковской сфере в ходе интервью при рассмотрении вопросов выдачи кредитов гражданам и т. д.

В статье представлен аналитический обзор компьютерных паралингвистических систем для автоматического распознавания лжи в речи человека, а также обобщенная схема автоматического паралингвистического анализа речи и классификация методов обработки аудиосигналов. Приведена информация о международных соревнованиях по компьютерной паралингвистике Computational Paralinguistic Challenge, представлена задача распознавания лжи в речи человека, которая ставилась в рамках последних соревнований, проходивших в США в 2016 г. Приведено описание речевой базы данных, критерии оценки систем, описание и анализ работ, представленных на конкурс.

Нерешенными на данный момент для задачи распознавания лжи в речи человека являются проблемы как технического характера (наличие аудиошумов, низкое качество сигнала в телефонном канале), так и естественного (высокая вариативность спонтанной речи, неоднозначность языка). Также важно учитывать индивидуальные особенности диктора, такие как физическое состояние, пол, возраст, эмоциональную стабильность/нестабильность, культурные и национальные отличия и т. д.

Данное исследование проводится при поддержке РФФИ (проекты № 16-37-60085 и 16-37-60100), Совета по грантам Президента РФ (гранты № МК-7925.2016.9 и МД-254.2017.8), а также бюджетной темы № 0073-2014-0005.

## Литература


1. **Карпов А. А., Кайа Х., Салах А. А.** Актуальные задачи и достижения систем паралингвистического анализа речи // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 4. С. 581–592. doi:10.17586/2226-1494-2016-16-4-581-592
2. **Горшков Ю. Г., Дорофеев А. В.** Речевые детекторы лжи коммерческого применения // ИНФОРМОСТ. «Радиоэлектроника и Телекоммуникации». 2003. № 6(30). С. 13–15.
3. **Montacié C., Caraty M.-J.** Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge // Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2016–2020.
4. **Будков В. Ю., Савельев А. И., Вольф Д. А.** Методика исследования параметров речевого сигнала, отражающая истинность передаваемой информации // Докл. ТУСУР. 2016. Т. 19. № 2. С. 56–60. doi:10.21293/1818-0442-2016-19-2-56-60
5. **Басов О. О., Карпов А. А., Саитов И. А.** Методологические основы синтеза полимодальных инфокоммуникационных систем государственного управления. — Орел: Академия ФСО РФ, 2015. — 271 с.
6. **Levitan S. I., An G., Ma M., Levitan R., Rosenberg A., Hirschberg J.** Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection// Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2006–2010.
7. **Herms R.** Prediction of Deception and Sincerity from Speech using Automatic Phone Recognition-based Features// Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2036–2040.
8. **Levitan S. I., An G., Wang M., Mendels G., Hirschberg J., Levine M., Rosenberg A.** Cross-Cultural Production and Detection of Deception from Speech// Proc. ACM Workshop on Multimodal Deception Detection, Seattle, USA, 2015. P. 1–8.
9. **Родькина О. Я., Никольская В. А.** К проблеме распознавания психоэмоционального состояния человека по речи с использованием автоматизированных систем // Информационные технологии. 2016. № 10(22). С. 728–733.
10. **Савченко В. В., Васильев Р. А.** Анализ эмоционального состояния диктора по голосу на основе фонетического детектора лжи // Научные ведомости Белгородского государственного университета. 2014. Вып. 32/1. № 21(192). С. 186–195.
11. **Amiriparian S., Pohjalainen J., Marchi E., Pugachevskiy S., Schuller B.** Is Deception Emotional? An Emotion-Driven Predictive Approach// Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2011–2015.

12. **Hirschberg J.** Detecting Deceptive Speech: Requirements, Resources and Evaluation// Proc. LREC-2008, Marrakech, Morocco, 2008. http://www.lrec-conf.org/proceedings/lrec2008/keynotes/Hirschberg.pdf (дата обращения: 30.03.2017).
13. **Kirchhubel C., Stedmon A., Howard D. M.** Analyzing Deceptive Speech// Proc. EPCE-2013. Springer LNCS. 2013. Vol. 8019. P. 134–141. doi:10.1007/978-3-642-39360-0_15
14. **Pan X., Zhao H., Zhou Y.** The Application of Fractional Mel Cepstral Coefficient in Deceptive Speech Detection // PeerJ. 2015. https://doi.org/10.7717/peerj.1194 (дата обращения: 30.03.2017). doi:10.7717/peerj.1194
15. **Schuller B., Steidl S., Batliner A., Hirschberg J., Burgoon J. K., Baird A., Elkins A., Zhang Y., Coutinho E., Evanini K.** The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language // Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2001–2005.
16. **Kaya H., Karpov A.** Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks// Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2046–2050.
17. **Zhang Y., Weninger F., Ren Z., Schuller B.** Sincerity and Deception in Speech: Two Sides of the Same Coin? A Transfer- and Multi-Task Learning Perspective// Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2041–2045.
18. **Lee H.-S., Tsao Y., Lee C.-C., Wang H.-M., Lin W.-C., Chen W.-C., Hsiao S.-W., Jeng S.-K.** Minimization of Regression and Ranking Losses with Shallow Neural Networks on Automatic Sincerity Evaluation// Proc. INTERSPEECH-2016, San Francisco, USA, 2016. P. 2031–2035.
19. **Levitan S. I., Levitan Y., An G., Levine M., Rosenberg A., Levitan R., Hirschberg J.** Identifying Individual Differences in Gender, Ethnicity, and Personality from Dialogue for Deception Detection// Proc. NAACL-HLT-2016, San Diego, USA, 2016. P. 40–44.
20. **Раисов М. Э., Мещеряков Р. В.** Полиграф на основе речевого ввода // Научная сессия ТУСУР-2009. Томск: В-Спектр, 2009. Ч. 3. С. 344–346.
21. Пат. 2293518 РФ. Способ оценки искренности/неискренности говорящего / Морозов В. П., Морозов П. В. — № 2005124844/14; заявл. 04.08.04; опубл. 20.02.07, Бюл. № 5. — 19 с.

UDC 621.391:004.934.2
doi:10.15217/issn1684-8853.2017.5.30

**Analytical Survey of Computational Paralinguistic Systems for Automatic Recognition of Deception in Human Speech**

Velichko A. N.[a,b], Master Student, velichko.a.n@mail.ru
Budkov V. Y.[a], PhD, Tech., budkov@iias.spb.su
Karpov A. A.[a], Dr. Sc., Tech., Associate Professor, karpov@iias.spb.su
[a]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg State University of Aerospace Instrumentation, 67, B. Morskaya St., 190000, Saint-Petersburg, Russian Federation

**Introduction:** Computational paralinguistics analyzes non-verbal aspects of human communication and speech, such as natural emotions, intonations, pronunciation features, speaker's voice parameters, truth of a message, etc. The problem of automatic detection of truth/deception in spoken messages has importance in many practical applications. There are a great number of contemporary studies devoted to the development of software for automated systems of human speech deception detection. **Purpose:** We analyze and discuss the achievements and developments in the field of computational paralinguistics, particularly deception detection in human speech, in order to figure out the drawbacks of the available methods and define the ways to overcome them in developing a new automatic system. **Results:** The analysis of a wide spectrum of state-of-the-art scientific and technical literature discussing the results of the world-wide scientific research in this field for the last ten years, including International Computational Paralinguistic Challenge (ComParE) has shown that the researchers apply similar methods for deception/truth detection. However, the signal processing algorithms have some differences which can affect the accuracy of the deception recognition. We present a generalized scheme of a recognition system, its main components, as well as a classification of the most efficient methods used in the development of automatic systems for paralinguistic analysis of natural speech. At present, human speech deception detection has a lot of unresolved problems, both of technical and natural types, including taking into account individual features of a speaker (gender, age, emotional stability, national specificity, etc.). Overcoming these problems can significantly improve the system functionality.
**Keywords** — Computational Paralinguistics, Speech Technologies, Human Speech Deception Detection, Machine Learning.

**References**

1. Karpov A. A., Kaya H., Salah A. A. State-of-the-Art Tasks and Achievements of Paralinguistic Speech Analysis Systems. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki* [Scientific and Technical Journal of Information Technologies, Mechanics and Optics], 2016, vol. 16, no. 4, pp. 581–592 (In Russian). doi:10.17586/2226-1494-2016-16-4-581-592
2. Gorshkov Y. G., Dorofeev A. V. Commercial Speech Deception Detectors. *INFORMOST. "Radioelektronika i Telekommunikatsii"*, 2003, no. 6(30), pp. 13–15 (In Russian).
3. Montacié C., Caraty M.-J. Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge. *Proc. INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 2016–2020.

4. Budkov V. Y., Savielev A. I., Volf D. A. Technique of Studying Speech Signal Parameters Reflecting on the Truth of the Transmitted Information. *Doklady TUSUR* [Proc. of TSUCSR], 2016, vol. 19, no. 2. pp. 56–60 (In Russian). doi:10.21293/1818-0442-2016-19-2-56-60
5. Basov O. O., Karpov A. A., Saitov I. A. *Metodologicheskie osnovy sinteza polimodal'nykh infokommunikatsionnykh sistem gosudarstvennogo upravleniya* [Methodological Basis for Synthesis of Polymodal Infocommunication Systems for State Administration]. Orel, Akademiia FSO RF Publ., 2015. 271 p. (In Russian).
6. Levitan S. I., An G., Ma M., Levitan R., Rosenberg A., Hirschberg J. Combining Acoustic-Prosodic, Lexical, and Phonotactic Features for Automatic Deception Detection. *Proc. INTERSPEECH-16*, San Francisco, USA, 2016, pp. 2006–2010.
7. Herms R. Prediction of Deception and Sincerity from Speech using Automatic Phone Recognition-based Features. *Proc. INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 2036–2040.
8. Levitan S. I., An G., Wang M., Mendels G., Hirschberg J., Levine M., Rosenberg A. Cross-Cultural Production and Detection of Deception from Speech. *Proc. ACM Workshop on Multimodal Deception Detection*, Seattle, USA, 2015, pp. 1–8.
9. Rodkina O. Ya., Nikolskaya V. A. To the Problem of Person's High Emotional State Recognition by his Speech. *Informatsionnye tekhnologii* [Information Technologies], 2016, no. 10(22), pp. 728–733 (In Russian).
10. Savchenko V. V., Vasilyev R. A. The Analysis of the Emotional Condition of the Announcer on the Voice on the Basis of the Phonetic Lie Detector. *Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta* [Belgorod State University Scientific Bulletin], 2014, iss. 32/1, no. 21(192), pp. 186–195 (In Russian).
11. Amiriparian S., Pohjalainen J., Marchi E., Pugachevskiy S., Schuller B. Is Deception Emotional? An Emotion-Driven Predictive Approach. *Proc. INTERSPEECH-2016*, San-Francisco, USA, 2016, pp. 2011–2015.
12. Hirschberg J. Detecting Deceptive Speech: Requirements, Resources and Evaluation. *Proc. LREC-2008*, Marrakech, Morocco, 2008. Available at: http://www.lrec-conf.org/proceedings/lrec2008/keynotes/Hirschberg.pdf (accessed 30 March 2017).
13. Kirchhubel C., Stedmon A., Howard D. M. Analyzing Deceptive Speech. *Proc. EPCE-2013*, Springer LNCS, 2013, vol. 8019, pp. 134–141. doi:10.1007/978-3-642-39360-0_15
14. Pan X., Zhao H., Zhou Y. The Application of Fractional Mel Cepstral Coefficient in Deceptive Speech Detection. *PeerJ*, 2015. Available at: https://doi.org/10.7717/peerj.1194 (accessed 30 March 2017). doi:10.7717/peerj.1194
15. Schuller B., Steidl S., Batliner A., Hirschberg J., Burgoon J. K., Baird A., Elkins A., Zhang Y., Coutinho E., Evanini K. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. *Proc. INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 2001–2005.
16. Kaya H., Karpov A. Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks. *Proc. INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 2046–2050.
17. Zhang Y., Weninger F., Ren Z., Schuller B. Sincerity and Deception in Speech: Two Sides of the Same Coin? A Transfer- and Multi-Task Learning Perspective. *Proc. INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 2041–2045.
18. Lee H.-S., Tsao Y., Lee C.-C., Wang H.-M., Lin W.-C., Chen W.-C., Hsiao S.-W., Jeng S.-K. Minimization of Regression and Ranking Losses with Shallow Neural Networks on Automatic Sincerity Evaluation. *Proc. INTERSPEECH-2016*, San Francisco, USA, 2016, pp. 2031–2035.
19. Levitan S. I., Levitan Y., An G., Levine M., Rosenberg A., Levitan R., Hirschberg J. Identifying Individual Differences in Gender, Ethnicity, and Personality from Dialogue for Deception Detection. *Proc. NAACL-HLT-2016*, San Diego, USA, 2016, pp. 40–44.
20. Raisov M. E., Meshcheryakov R. V. Polygraph Based on Speech Input. *Nauchnaia sessiia TUSUR-2009*, Tomsk, vol. 3, pp. 344–346 (In Russian).
21. Morozov V. P., Morozov P. V. *Sposob otsenki iskrennosti/neiskrennosti govoriashchego* [A Method for Estimating Sincerity-Insincerity of a Speaker]. Patent RU, no. 2293518, 2007.

## УВАЖАЕМЫЕ АВТОРЫ!

Научные базы данных, включая SCOPUS и Web of Science, обрабатывают данные автоматически. С одной стороны, это ускоряет процесс обработки данных, с другой — различия в транслитерации ФИО, неточные данные о месте работы, области научного знания и т. д. приводят к тому, что в базах оказывается несколько авторских страниц для одного и того же человека. В результате для всех по отдельности считаются индексы цитирования, снижая рейтинг ученого.

Для идентификации авторов в сетях Thomson Reuters проводит регистрацию с присвоением уникального индекса (ID) для каждого из авторов научных публикаций.

Процедура получения ID бесплатна и очень проста: входите на страницу http://www.researcherid.com, слева под надписью «New to ResearcherID?» нажимаете на синюю кнопку «Join Now It's Free» и заполняете короткую анкету. По указанному электронному адресу получаете сообщение с предложением по ссылке заполнить полную регистрационную форму на ORCID. Получаете ID.