МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ИЗВЕСТИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ
ПРИБОРОСТРОЕНИЕ
ИЗДАНИЕ САНКТ-ПЕТЕРБУРГСКОГО НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО УНИВЕРСИТЕТА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
Журнал издается с января 1958 г.

ТОМ 57 **ФЕВРАЛЬ 2014** **№ 2**

ТЕМАТИЧЕСКИЙ ВЫПУСК

РЕЧЕВЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

Под редакцией кандидата технических наук М. В. Хитрова, доктора технических наук, профессора Ю. Н. Матвеева

СОДЕРЖАНИЕ

СИСТЕМЫ РАСПОЗНАВАНИЯ ЛИЧНОСТЕЙ ПО ГОЛОСУ

THEMATIC ISSUE

SPEECH INFORMATION SYSTEMS

By Edition of M. V. Khitrov, Candidate of Technical Science
Yu. N. Matveev, Doctor of Technical Science, Professor

CONTENTS

ПРЕДИСЛОВИЕ

Кафедра речевых информационных систем создана в 2011 г. на факультете Информационных технологий и программирования (ФИТиП) НИУ ИТМО. Она ориентирована на подготовку специалистов, способных участвовать в исследовательской и проектной работе в области речевых информационных технологий со специализацией в направлениях распознавания и синтеза речи, распознавания личности по голосу, мультимодальной биометрии, в области проектирования и разработки информационных систем и программного обеспечения.

Организатором создания кафедры выступила компания ООО „Центр речевых технологий“ (ЦРТ). Компания ЦРТ была создана в 1990 г. в Санкт-Петербурге и за 20 лет стала абсолютным лидером российского и значимым игроком международного рынка речевых технологий. Компания является ведущим мировым разработчиком систем в сфере высококачественной записи, обработки, анализа, синтеза и распознавания речи, голосовых и мультимодальных биометрических систем.

ЦРТ сегодня — активный участник быстрорастущего мирового рынка речевых и биометрических технологий. Компания поставляет свои решения более чем в 65 стран мира и ярко заявляет о себе в области инноваций — не только создает и внедряет уникальные разработки в сфере речевых технологий, но и фактически формирует новые сегменты рынка.

К преподаванию, а также проведению научно-исследовательских и опытно-конструкторских работ на кафедре речевых информационных систем привлекаются ведущие специалисты ЦРТ, преподаватели НИУ ИТМО, а также сотрудники других научных и коммерческих организаций.

В настоящем сборнике представлены результаты научно-исследовательских работ, выполняемых на кафедре речевых информационных систем НИУ ИТМО.

Заведующий кафедрой
речевых информационных систем НИУ ИТМО,
генеральный директор
ООО „ЦРТ“,
канд. техн. наук М. В. ХИТРОВ

Профессор кафедры
речевых информационных систем НИУ ИТМО,
главный научный сотрудник
ООО „ЦРТ-инновации“,
докт. техн. наук Ю. Н. МАТВЕЕВ

PREFACE

The Department of Speech Information Systems was established in 2011 at the Faculty of Information Technologies and Programming of St. Petersburg National Research University of Information Technologies, Mechanics and Optics (NRU ITMO). The Department was involved in training of specialists able to collaborate in R&D in the field of speech information technologies with special focus on speech recognition and synthesis, person identification by voice, multi-modal biometry, as well as on research and development of information systems and software.

The initiative in the Department establishing belonged to the company Speech Technology Center Ltd. (STC) created in St. Petersburg in 1990. In 20 years the company has become the absolute leader in Russian and a considerable player in international market of speech technologies. The company is a world-wide leader in development of systems for high-quality record, processing, analysis, synthesis, and recognition of speech, of voice and multi-modal biometric systems.

Today STC is an active participant of the quick-growing world market of speech and biometric technologies. The company delivers its solutions to more than 65 countries and makes itself known in the field of innovations — STC not only creates and implements unique products related to speech technologies, but also practically founds new segments of the market.

Teaching and organization of scientific investigations and development works at the Department of Speech Information Systems is performed with participation from leading specialists of STC, lecturers of NRU ITMO, as well as employees of other scientific and commercial institutions.

This issue presents results of scientific researches carried out at the Department of Speech Information Systems NRU ITMO.

Head of the Department of Speech Information Systems, NRU ITMO
General Director, Speech Technology Center Ltd.
Cand. Techn. Sci.
M. V. KHITROV

Professor, Department of Speech Information Systems, NRU ITMO
Chief Researcher, STC-Innovation Ltd.
Dr. Techn Sci.
Yu. N. MATVEEV

СИСТЕМЫ АНАЛИЗА РЕЧИ

УДК 57.087.1

М. В. Хитров, А. Ю. Васильев

СТАТИСТИЧЕСКИЕ ЯЗЫКОВЫЕ ОСОБЕННОСТИ И ИХ ГЕНДЕРНЫЕ РАЗЛИЧИЯ НА ПРИМЕРЕ ЛИТОВСКОГО ЯЗЫКА

Выявлены речевые особенности, позволяющие решать задачи автоматической идентификации языка и идентификации диктора. Предложен метод, использующий статистические параметры, характеризующие мелодический контур фраз исследуемого языка.

Ключевые слова: *речевые технологии, статистические аспекты языка, литовский язык.*

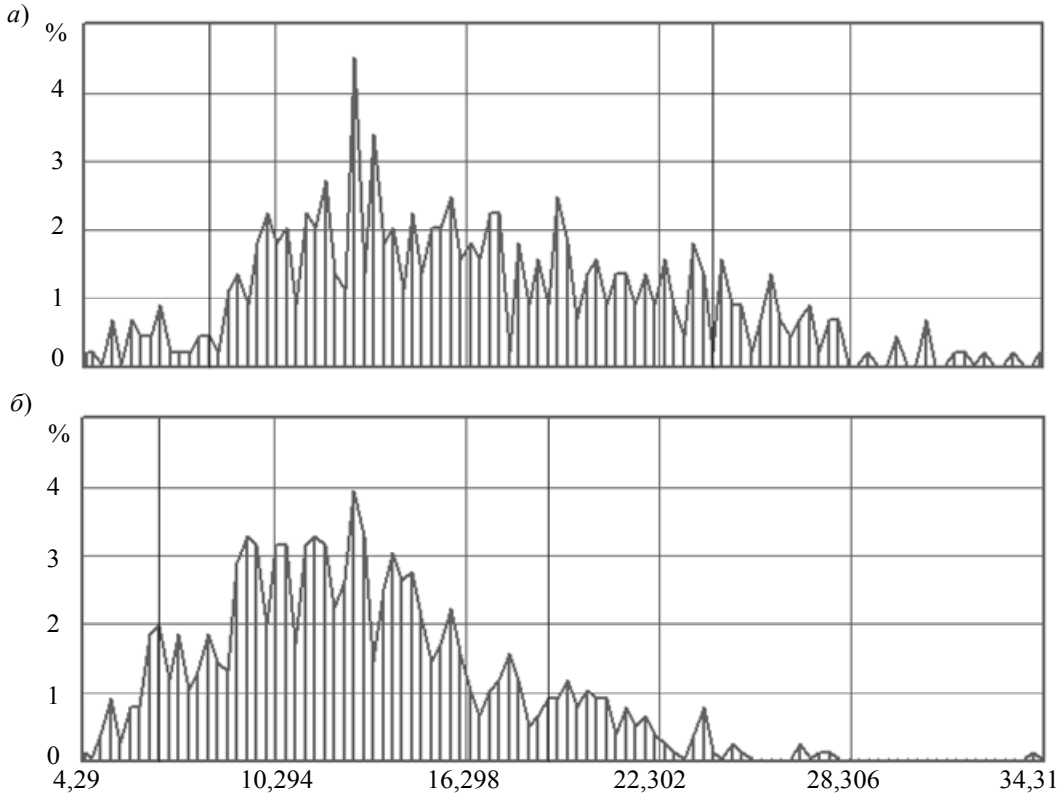
Введение. Определение речевых особенностей языка диктора — одна из актуальных задач современных речевых технологий. Одним из методов ее решения является сопоставление статистических характеристик основного тона говорящего и выявление типичных для выбранного языка интервалов статистических параметров различных речевых фрагментов [1—3]. Важно отметить, что границы данных интервалов и частота появления определенных речевых фрагментов существенно зависят от пола говорящего. В настоящей работе проанализированы статистические особенности языка и гендерные различия характеристик и речевых фрагментов на основе литовского языка.

База исследований. Исследование проводилось на 78 фонограммах литовских дикторов (на 55 фонограммах представлены 15 мужчин, на 23 — 7 женщин). Фонограммы содержали как литовскую, так и русскую речь. Обработка материала включала получение файлов основного тона, его минимальную коррекцию, а также подготовку таблиц, содержащих числовые значения параметров характерных интонационных единиц речи последовательностей фраз (длительность речевого абзаца 10—20 с). В качестве характерных интонационных единиц использовались синтагмы и их структурные элементы (ядерные слоги и ядерно-заядерные участки, шкалы, предшкалы [4—6]). Для анализа выбирались фрагменты речи, содержащие утвердительные высказывания, имеющие характер завершенности или незавершенности и одинаковую эмоциональную окраску.

В качестве параметров рассматриваемых речевых абзацев и локальных фрагментов выступали максимальная, минимальная и средняя частоты фрагмента (в герцах), частотный интервал фрагмента (в герцах и полутонах), скорость изменения тона (в полутонах в секунду), коэффициент изрезанности фрагмента [1, 2].

Анализ статистических характеристик. Значения каждого параметра для всех исследуемых речевых абзацев всех дикторов одного пола строятся в виде диаграмм. Специально разработанное программное обеспечение позволяет определять для полученного распределения основные статистические характеристики, включая математическое ожидание, среднеквадратичное отклонение, а в качестве конечного результата — интервал типичных значений с доверительной вероятностью 95 %.

Ниже приводятся два примера таких распределений параметра „интервал в полутонах“ для случаев литовских дикторов-женщин (рисунок, *а*) и мужчин (*б*). На рисунке отчетливо видны установленные границы доверительного интервала, соответствующие диапазону типичных значений характеристик для литовского языка.



Анализ распределений параметров показал, что наиболее близкий к нормальному характер, как у женщин, так и у мужчин, имеют изрезанность, максимальная частота и скорость изменения тона. Значения интервалов, минимальной и средней частот имеют распределение, промежуточное между нормальным и равномерным. Стоит отметить ряд аномальных пиков как вне, так и в области доверительного интервала, для женщин и средней частоты у мужчин, а также существенно выходящие за пределы доверительного интервала значения ряда статистических характеристик у мужчин (см. рисунок, *б*). Это может быть объяснено дефектом записанной фонограммы, процент таких отклонений входит в допустимую статистическую погрешность (табл. 1—4).

Таблица 1

Интервалы типичных значений параметров для литовского языка (мужчины)

Параметр	Область значений (доверительный интервал 95 %)
Максимум, Гц	129,9—217,4
Минимум, Гц	63,9—103,6
Интервал-Гц	41,7—136,8
Интервал-Пт	6,7—18,8
Средняя частота, Гц	90,1—143,7
Скорость изменения тона	–9,1—1,3
Эксцесс	–0,9—1,7
Скос	–0,15—1,68
Изрезанность	24,3—49,3

Таблица 2

Интервалы типичных значений параметров для тюркских языков (мужчины)

Параметр	Область значений (доверительный интервал 95 %)
Максимум, Гц	133,6—218,8
Минимум, Гц	66,3—97,4
Интервал-Гц	49,8—131,6
Интервал-Пт	8,3—17,7
Средняя частота, Гц	95—141,9
Скорость изменения тона	–12,1— –0,07
Эксцесс	2,3—4,6
Скос	–0,17—1,47
Изрезанность	27,5—50,25

Таблица 3

Интервалы типичных значений параметров для литовского языка (женщины)

Параметр	Область значений (доверительный интервал 95 %)
Максимум, Гц	225,8—395,7
Минимум, Гц	102,9—164,6
Интервал-Гц	88,4—254,5
Интервал-Пт	8,7—19,6
Средняя частота, Гц	156,3—242
Скорость изменения тона	–7,3—1,5
Эксцесс	–0,8—2,8
Скос	0,03—1,9
Изрезанность	26,6—46,4

Таблица 4

Интервалы типичных значений параметров для тюркских языков (женщины)

Параметр	Область значений (доверительный интервал 95 %)
Максимум, Гц	255,1—383,4
Минимум, Гц	124,4—196,4
Интервал-Гц	96,3—221,7
Интервал-Пт	8,2—16,3
Средняя частота, Гц	177,7—251,8
Скорость изменения тона	–9,3—0,52
Эксцесс	2,5—5,3
Скос	0,09—1,6
Изрезанность	28,3—45,2

Определенные интервальные значения явно демонстрируют гендерные различия: речевые фрагменты женского голоса имеют тенденцию к более высоким частотам, чем у мужского голоса. Также для речи женщин характерен более широкий диапазон частот. Прочие статистические характеристики в целом совпадают для обоих полов, их особенности проявляются при сравнении с другими языками.

Для сравнительной оценки типичных значений параметров литовского языка использованы усредненные типичные значения параметров узбекского и азербайджанского языков. Интервалы типичных значений для обоих тюркских языков получены на материале речевых абзацев большой длительности, полученном от 59 носителей азербайджанского языка (13 женщин, 46 мужчин) и 84 носителей узбекского языка (49 женщин, 35 мужчин). Сравнение проводилось отдельно для мужчин и для женщин.

В результате установлено, что для мужчин-литовцев по сравнению с тюркоязычными дикторами минимальные значения частоты основного тона (ЧОТ) варьируют в более широком диапазоне, максимальные значения ЧОТ также более вариативны, но диапазон их варьирования лежит в низкочастотной области. Частотный диапазон (выраженный в герцах и в полутонах) у литовцев шире, чем у узбеков и азербайджанцев. Коэффициент изрезанности для литовцев ниже, чем для носителей узбекского и азербайджанского языков, что указывает на более ровную интонацию во фразах литовцев.

Для литовских женщин-дикторов максимальные значения ЧОТ варьируют в более широком диапазоне по сравнению с тюркоязычными женщинами, а минимальные лежат в низкочастотной области и варьируют в более узких пределах. Судя по значениям параметра „средняя частота“, голоса литовских женщин в целом ниже, чем голоса тюркоязычных. Скорость изменения тона у литовских женщин выше, чем у узбекских и азербайджанских; значение коэффициента изрезанности у них варьирует в более широком диапазоне.

Кроме сравнения статистических параметров литовского языка со значениями для конкретной языковой группы — тюркской — было проведено исследование общих усредненных параметров по всем имеющимся данным для различных языковых групп. Результаты показали следующие особенности. Максимальные значения характерных частот литовского языка ниже средних характеристик по всем языкам базы, особенно ярко это выражено у мужчин, поэтому в связи с тем, что значения минимальных частот в целом соответствуют усредненным данным, характерный частотный интервал у литовцев меньше (у мужчин — примерно на четверть). Средняя скорость изменения тона для литовского языка невысока, у мужчин отклонения от средних значений достигают 50 %. Изрезанность характерных речевых фрагментов свидетельствует о более плавных, в сравнении со средними значениями, изменениях частоты основного тона. В целом литовскому языку присуща плавность речи в сравнительно узком частотном диапазоне. Для речи дикторов-мужчин типичны существенные отклонения от усредненных характеристик по всей речевой базе, женщины демонстрируют большее соответствие. Поскольку база содержит характеристики различных языковых групп, трудно делать выводы о близости характеристик для литовского языка и родственных ему языков балтийской ветви.

Анализ речевых фрагментов. Анализ набора интонационных структур показал, что отличительной особенностью речи женщин является более частое использование восходящего мелодического контура.

Так, в приведенном материале (табл. 5) в речи мужчин на русском и литовском языках встретилось одинаковое количество реализаций синтагм с восходящим и нисходящим завершением, в то время как в речи женщин синтагмы с восходящим завершением реализуются значительно чаще, чем с нисходящим (363 и 189 реализаций соответственно).

Таблица 5

Тип	Подтип	Фрагменты речи		Дикторы	
		мужчины	женщины	мужчины	женщины
Синтагма	Нисходящее завершение	598	189	15	7
	Восходящее завершение	598	**363**	15	7
Ядерно-заядерный участок	Нисходящий	**475**	142	15	7
	Восходящий	415	**195**	15	7
	Нисходяще-восходящий	83	**101**	12	7
	Восходяще-нисходящий	58	18	13	6

Та же закономерность наблюдается в реализации ядерно-заядерных участков синтагм в речи литовских дикторов. Мужчины наиболее часто используют нисходящую интонацию ядерно-заядерного участка, восходящая интонация используется реже. Нисходяще-восходящее и восходяще-нисходящее оформление встречается значительно реже. Женщины чаще реализуют восходящий контур ядерно-заядерного участка, нисходящая интонация используется ими реже. В речи женщин значительно чаще, чем у мужчин, реализуется нисходяще-восходящая интонация ядерно-заядерного участка.

Заключение. По результатам проведенных исследований можно сделать следующие выводы. Литовский язык по своим речевым особенностям тяготеет к более низким частотам, что особенно выражено в речи мужчин. Для него также характерны достаточно узкий частотный диапазон и плавность изменения частоты основного тона. Женская и мужская речь в литовском языке различается не только очевидным тяготением мужских голосов к низким частотам, а женских — к высоким, но и тем, что статистические параметры мужского голоса сильнее отличаются от средних для различных языковых групп, а также тем, что женщины демонстрируют стремление к интонации незавершенности в утвердительных высказываниях. Сделанные выводы справедливы для утвердительных высказываний одинаковой эмоциональной окраски.

В настоящей работе впервые представлены возможности использования метода статистического анализа мелодического контура речи в различных задачах, включая задачи идентификации языка и диктора. Логичное продолжение исследований — сбор и анализ данных не только по отдельным языкам, но и по значимым языковым группам, поскольку это должно повысить точность результатов и послужить первым шагом в разработке инструментальных программных средств анализа особенностей языка. Точность анализа и идентификации параметров будет различаться для языка внутри своей языковой группы и языков разных групп. Другим направлением может служить анализ речевых фрагментов различной эмоциональной окраски.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Коваль С. Л., Лабутин П. В., Пеховский Т. С., Прощина Е. А., Смирнова Н. А., Таланов А. О.* Методика идентификации дикторов по голосу и речи на основе комплексного анализа фонограмм // Тр. Междунар. конф. „Диалог 2007“. М., 2007. С. 256—262.
2. *Смирнова Н. А.* Идентификация дикторов на основе сравнения параметров реализации мелодических контуров высказываний // Тр. Междунар. конф. „Диалог 2007“. М., 2007. С. 502—507.
3. *Хитров М. В.* и др. Фоноскопическое исследование фонограмм речи: Исследование достоверности фонограмм. СПб: Изд-во „Юридический центр-Пресс“, 2011. Кн. I. 281 с.
4. *O'Connor J., Arnold G.* Intonation of colloquial English. London: Longman, 1973.
5. *Брызгунова Е. А.* Интонация // Русская грамматика / *Н. Ю. Шведова* (гл. ред.). М.: Наука, 1980. С. 96—122.
6. *Светозарова Н. Д.* Интонационная система русского языка. Л.: Изд-во ЛГУ, 1982.

Сведения об авторах

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

Андрей Юрьевич Васильев — ООО „ЦРТ“, Санкт-Петербург; программист; E-mail: vasilyev-a@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 519.86

В. В. Киселёв, А. В. Ткаченя, М. В. Хитров

РАЗРАБОТКА КАНАЛОНЕЗАВИСИМЫХ ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Исследованы информативные признаки речи с целью формирования каналонезависимого пространства признаков для повышения эффективности функционирования системы распознавания дикторов. Экспериментально определен оптимальный набор каналонезависимых информативных признаков для решения задачи выявления сходства между фонограммами на основе метода динамического программирования.

Ключевые слова: *голосовой анализ, машинное обучение, выбор информативных признаков, мел-частотные кепстральные коэффициенты, метод динамического программирования.*

Введение. Важнейшим этапом в создании систем автоматического голосового анализа является выделение оптимального набора информативных признаков. При решении большинства прикладных задач анализу подвергаются голосовые данные диктора, полученные при различных условиях записи. Изменение характеристик канала приводит к изменению анализируемого пространства признаков, что снижает эффективность классификации дикторов.

Цель предлагаемой работы — снижение влияния характеристик канала на эффективность работы систем голосового анализа. Для достижения цели необходимо использовать каналонезависимые информативные признаки. В последнее время исследования в этом направлении приобрели особую актуальность [1—3]. Тем не менее, большинство существующих способов получения каналонезависимых информативных признаков характеризуются большими временными и аппаратными затратами, что затрудняет их использование в задачах, требующих анализа сигнала в реальном масштабе времени.

В настоящей работе сравнивается эффективность для случая использования исходных информативных и полученных каналонезависимых признаков на примере задачи выявления сходства между фонограммами. Для этого применяется метод динамического программирования (DTW), заключающийся в последовательном сравнении анализируемой записи с образцом. При помощи DTW происходит сравнение массивов информативных признаков анализируемой записи и образца произношения. Данный подход часто используется при построении простых систем распознавания речи [4, 5].

Алгоритм сравнения фонограмм. Анализ фонограмм выполняется в соответствии с блок-схемой, приведенной на рис. 1.

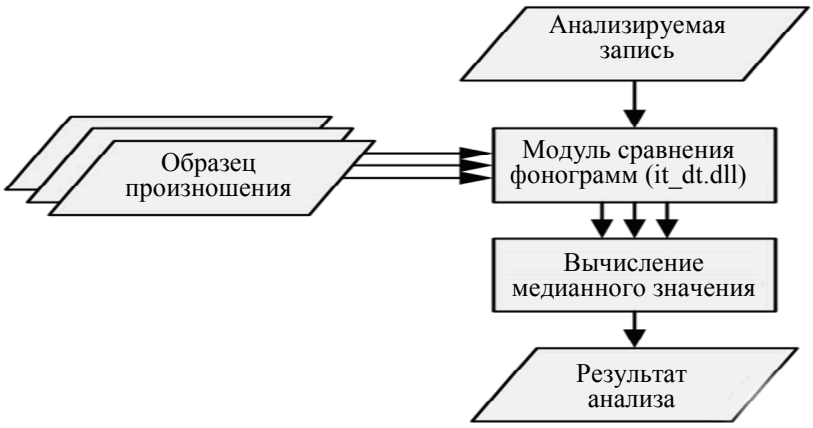


Рис. 1

Из рисунка видно, что анализируемая запись сравнивается с каждым из образцов правильного произношения, а конечный результат анализа вычисляется как медианное значение результатов сравнения отдельных фонограмм. Использование медианного значения позволяет получить устойчивую оценку степени сходства фонограмм и обусловлено необходимостью исключения чрезмерной адаптации к конкретному образцу произношения.

Сравнение каждой фонограммы-образца произношения с анализируемой записью выполняется в соответствии со схемой, приведенной на рис. 2.

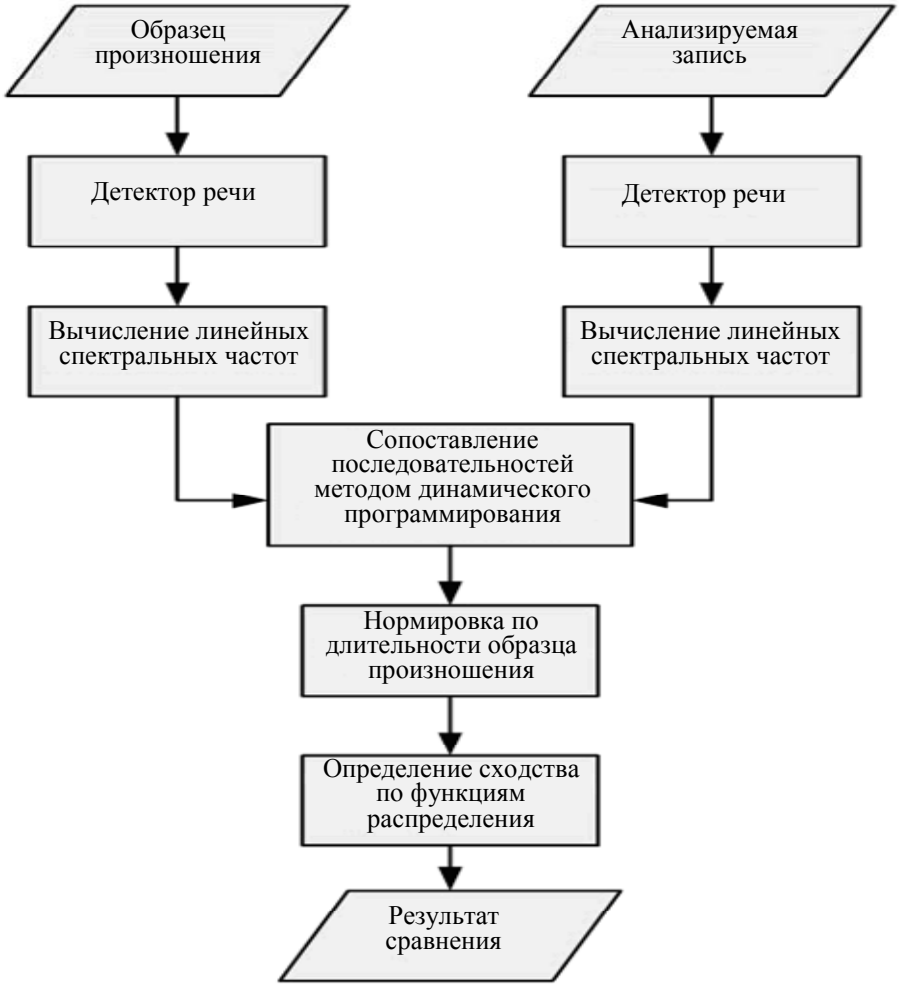


Рис. 2

Особенность предложенного алгоритма сравнения двух фонограмм заключается в использовании блока нормирования по длительности образца произношения, что позволяет снизить временные и аппаратные затраты на сопоставление анализируемой записи с образцом.

Выбор информативных признаков. Известно, что чувствительность человека к звуковому сигналу зависит от частоты сигнала: чем ниже частота, тем чувствительность выше. В 1937 г. была выведена формула, по которой можно перевести частоту (f) в герцах в частоту в мелах (m):

$$m = 1127,01048\ln(1+f/700), \quad f = 700(e^{m/1127,01048}-1).$$

Сигнал представляется как свертка двух функций: исходного сигнала и фильтра, параметры которого должны быть оценены. Необходимо разделить эти отдельные компоненты при помощи преобразования

$$x * h = \hat{x} + \hat{h}\,.$$

Для этого вводится кепстральное преобразование
— вещественный кепстральный коэффициент:

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|X(e^{i\omega})\right| e^{i\omega n} d\omega\,;$$

— комплексный кепстральный коэффициент:

$$C[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(X(e^{i\omega})) e^{i\omega n} d\omega\,,$$

где $X(e^{i\omega})$ — спектр сигнала; ω — частота (в радианах).

Такой подход позволяет получить характеристики речевого сигнала (мел-частотные кепстральные коэффициенты, MFCC), которые минимально зависят от индивидуальных особенностей говорящего, а значит, могут быть очень полезны в задачах распознавания [6].

Так как при решении прикладных задач анализируются данные, полученные в различных условиях записи, изменяется анализируемое пространство признаков и снижается эффективность классификации. Для достижения робастности голосового анализа в системах распознавания диктора необходимо использовать каналонезависимые информативные признаки.

Часто в литературе нормировка параметров канала связи (адаптация коэффициентов наблюдений) выполняется посредством вычитания средних значений коэффициентов вещественного кепстра. Такой подход позволяет эффективно бороться с мультипликативными искажениями, вносимыми различными каналами связи.

Вычитание средних значений MFCC вместо вычитания средних значений коэффициентов вещественного кепстра накладывает определенные ограничения на виды допустимых мультипликативных искажений, однако более эффективно в вычислительном плане.

Возможны различные способы оценки среднего значения мел-кепстральных коэффициентов:

1) оценка средних значений на неречевых участках, этот способ позволяет эффективно бороться с мультипликативными искажениями канала связи, сохраняя информацию об индивидуальных голосовых характеристиках диктора;

2) оценка средних значений как на вокализованных, так и на невокализованных участках речи;

3) оценка средних значений только на вокализованных участках речи, что позволяет нормировать коэффициенты наблюдений как к каналу связи, так и к голосу диктора. За счет того, что средние значения оцениваются только на вокализованных участках речи, дисперсии оценок оказываются меньше, чем при оценке средних на вокализованных и невокализованных участках речи.

При необходимости работы в режиме реального времени для вычитания среднего часто применяется фильтр с коэффициентами **b** = [1 –1], **a**=[1 –0,97]. При этом инициализация фильтра выполняется таким образом, чтобы x_0=x_1, y_0=0. АЧХ (*2*) и ФЧХ (*1*) такого фильтра приведены на рис. 3 ($\overline{f} = f\pi$ радиан/отсчет).

Для того чтобы информативные признаки стали каналонезависимыми, было предложено провести оценку средних значений только на вокализованных участках речи. Такой шаг позволяет вышеописанные мел-частотные кепстральные коэффициенты, сильно зависящие от

характеристик канала, сделать каналонезависимыми и значительно повысить эффективность использующих их систем.

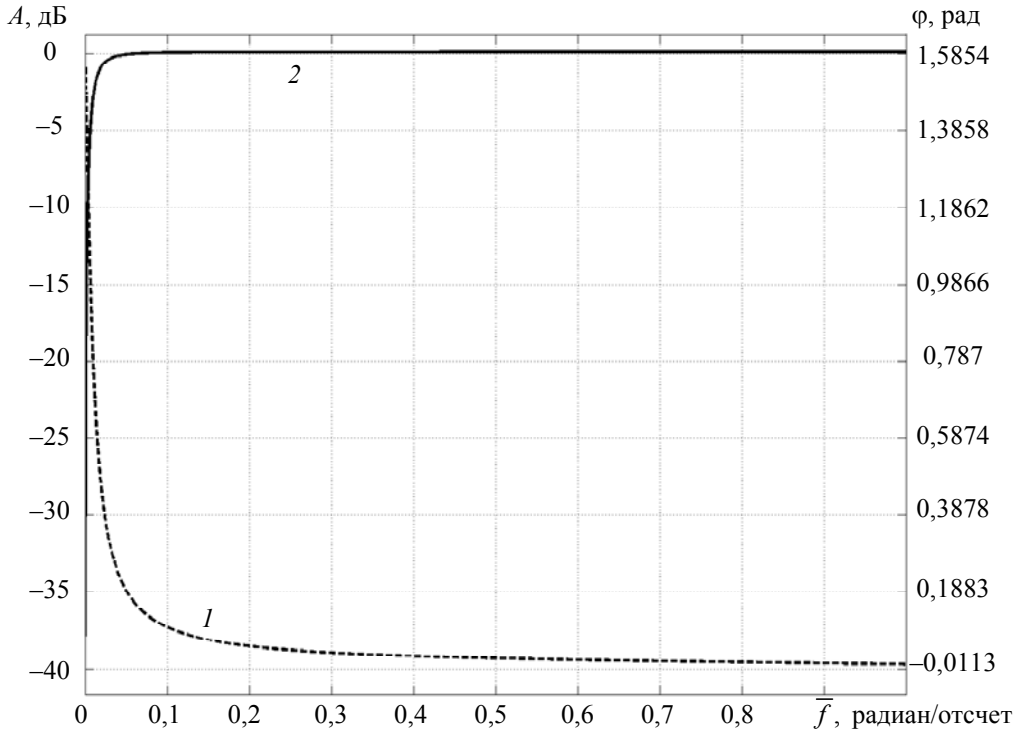


Рис. 3

Сравнение фонограмм. Ранее последовательности MFCC сопоставлялись методом динамического программирования [6]. DTW позволяет найти максимальное подобие двух заданных последовательностей, при этом мера их подобия не зависит от изменения нелинейного масштаба времени. Эти свойства DWT наилучшим образом подходят для решения поставленной задачи сравнения фонограмм.

С целью формирования матрицы локальных расстояний d_{ij} для каждой пары сравниваемых MFCC вычисляется L1-метрика:

$$d_{ij} = \sum_{n=1}^{p} \left| \mathrm{MFCC}_{in} - \mathrm{MFCC}_{jn} \right|.$$

Определение матрицы интегральных расстояний D_{ij} выполняется с использованием локальных ограничений Итакуры [7]:

$$D_{ij} = \min \begin{Bmatrix} D_{i-2,j-1} + d_{i-1,j} \\ D_{i-1,j-1} \\ D_{i-1,j-2} + d_{i,j-1} \end{Bmatrix}.$$

Расстоянием между сравниваемыми записями является значение матрицы интегральных расстояний с максимальными значениями индексов $D_{\max\_i,\max\_j}$.

Результаты экспериментов. Разработанный алгоритм сравнения фонограмм предназначен для контроля правильности произношения слов и выражений при обучении языкам. Работа алгоритма предусматривает запись пользователем требуемой речевой фонограммы и получение комплексной оценки меры подобия записанного сигнала с заданными образцами произношения (см. рис. 1).

Для проведения эксперимента были выбраны три типа фонограмм: одиночное слово, фраза (до 7 слов) и скороговорка. В тестировании принимали участие 4 диктора (2 мужчины и 2 женщины), не вошедшие в обучающую выборку. Проверка эффективности работы алгоритма оценки сходства фонограмм проводилась на файлах, записанных при следующих условиях: соотношение сигнал/шум (SNR) 15 и 30 дБ, клиппирование сигнала (clipping) [8], одиночная ошибка (1 miss), множественная ошибка (N miss). Результаты тестирования приведены в таблице.

Степень сходства анализируемых записей при различных шумах и искажениях

Информативный признак	SNR 15 дБ	SNR 30 дБ	Clipping	1 miss	N miss
1 слово					
MFCC	57	92	46	75	42
Каналонезависимые MFCC	79	93	68	77	44
Фраза					
MFCC	54	88	37	80	45
Каналонезависимые MFCC	76	90	60	79	40
Скороговорка					
MFCC	53	89	38	83	49
Каналонезависимые MFCC	74	91	63	80	42

Заключение. В статье предложен метод формирования каналонезависимого пространства признаков классификатора на основе MFCC. Было проведено экспериментальное исследование эффективности предложенного метода, включающее определение оптимального набора параметров и построение классификатора для выявления сходства фонограмм. Такой способ построения каналонезависимых информативных признаков характеризуется низкими временными и аппаратными затратами, что позволяет их использовать в системах голосового анализа без значительного снижения производительности конечного программного комплекса.

Как видно из таблицы, использование каналонезависимых информативных признаков приводит к повышению точности разделения правильного и неправильного произношения фонограммы. При этом эффективность классификации зашумленных и клиппированных сигналов значительно возросла: в среднем на 20—25 %.

В качестве дальнейшей работы представляется целесообразным протестировать эффективность применения описанных каналонезависимых информативных признаков для определения психоэмоционального состояния человека по его речи.

СПИСОК ЛИТЕРАТУРЫ

1. *Moritz N., Anemüller J., Kollmeier B.* Amplitude Modulation Filters as Feature Sets for Robust ASR: Constant Absolute or Relative Bandwidth? // Proc. 13th Annual Conf. of the Intern. Speech Communication Association (Interspeech-2012). Portland, Oregon, USA, 2012. P. 1230—1233.

2. *Meyer B. T., Spille C., Kollmeier B., Morgan N.* Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition // Proc. 13th Annual Conference of the International Speech Communication Association (Interspeech-2012). Portland, Oregon, USA, 2012. P. 1258—1261.

3. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 47—51.

4. *Kraljevski I., Gacovski Z., Arsenovski S., Mihajlov M.* Performance of DTW Speech Recognizer on Packet Switched Network // Proc. VII ETAI Conf. Ohrid, Macedonia, 2005. P. 16—20.

5. *Paliwal K. K.* On the Use of line Spectral Frequency Parameters for Speech Recognition // Digital Signal Processing. 1992. Vol. 2. P. 80—87.

6. *Rabiner L., Biing-Hwang Juang*. Fundamentals of speech recognition. Inc. Upper Saddle River, NJ, USA: Prentice-Hall, 1993. 496 p.

7. *Keogh E., Ratanamahatana C.A.* Exact indexing of dynamic time warping // Knowledge and Information Systems. 2005. Vol. 7, Is. 3. P. 358—386.

8. *Алейник С. В., Матвеев Ю. Н., Раев А. Н.* Метод оценки уровня клиппирования речевого сигнала // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 3 (79). С. 79—83.

Сведения об авторах

Виталий Владимирович Киселёв — ООО „Речевые технологии“, Минск; директор; E-mail: kiselev-v@speechpro.com

Андрей Владимирович Ткаченя — ООО „Речевые технологии“, Минск; младший научный сотрудник; E-mail: tkachenia-a@speechpro.com

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 004.934

Н. А. Томашенко, Ю. Ю. Хохлов

ИССЛЕДОВАНИЕ ПРОБЛЕМЫ СБАЛАНСИРОВАННОСТИ ДАННЫХ ПРИ ПОСТРОЕНИИ АКУСТИЧЕСКИХ МОДЕЛЕЙ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Исследована проблема сбалансированности данных при обучении акустических моделей для автоматического распознавания речи. Предложена метрика, позволяющая при кластеризации состояний трифонов явно учитывать влияние количества данных в кластере. Экспериментально доказано, что использование такого подхода позволяет повысить качество распознавания речи.

Ключевые слова: *автоматическое распознавание речи, GMM-HMM, обучение акустических моделей, связывание состояний, сбалансированность данных, кластеризация, трифоны.*

Введение. Качество системы автоматического распознавания речи в значительной степени определяется характеристиками используемых в ней акустических моделей. В настоящее время в области распознавания речи обычно применяются статистические подходы, при этом свойства акустических моделей во многом зависят от характеристик речевой базы данных, на которой эти модели были обучены. Одна из наиболее распространенных проблем, связанных с речевыми базами данных, — различие объемов (несбалансированность) данных, приходящихся на разные акустические классы, что может оказывать серьезное влияние на классифицирующую способность моделей [1]. В частности, отсутствие необходимого количества данных в обучающей выборке для определенных моделей усложняет получение надежной оценки параметров этих моделей.

Проблеме несбалансированности классов уделено много внимания в литературе по машинному обучению (см., например, [2]). Несмотря на то что многие алгоритмы обучения предполагают сбалансированность данных, это условие не всегда выполняется для реальных приложений, когда одни классы представлены большим количеством данных в обучающей выборке, а другие — всего несколькими элементами. Этой особенностью отличаются и речевые базы данных, используемые при построении акустических моделей.

Методы решения задачи, связанной с несбалансированностью данных, можно разделить на те, которые направлены на модификацию обучающей выборки и ее балансировку [1, 3], и те, которые преобразовывают сам алгоритм обучения (к последним относится метод, предложенный в настоящей работе).

Цель настоящей работы — исследование влияния сбалансированности данных на качество распознавания. Предметом исследования являются скрытые марковские модели (Hidden Markov Models, HMM), в которых состояния моделей фонем или контекстных трифонов (фонем с определенным левым и правым контекстом) описываются с помощью смеси гауссовых распределений (Gaussian Mixture Models, GMM).

Связывание состояний моделей трифонов при обучении акустических моделей является центральным механизмом в регулировании соотношения между сложностью модели (количеством параметров) и количеством данных в обучающей выборке. Редкие трифоны на уровне состояний или моделей [4] связываются с другими трифонами.

Многие алгоритмы обучения акустических моделей (кластеризации трифонов) [4—10] — агломеративный data-driven (управляемый данными), tree-based clustering (кластеризация на основе фонетического дерева) или их модификации — не позволяют напрямую задавать степень влияния количества данных на параметры связывания состояний трифонов. В настоящей работе предложена метрика, позволяющая по количеству данных в процессе кластеризации учитывать сбалансированность классов, а также исследована зависимость качества акустических моделей от степени этого влияния.

Кластеризация трифонов. Существует два основных метода кластеризации состояний трифонов при обучении акустических моделей на основе GMM-HMM.

1. Агломеративная кластеризация. Этот метод представляет собой „восходящую“ (bottom-up) процедуру кластеризации. Изначально все состояния трифонов рассматриваются как отдельные кластеры. Далее выбирается пара кластеров, при объединении образующих наименьший кластер. Процесс объединения кластеров продолжается до тех пор, пока размер самого большего кластера не достигнет заданного порога либо пока число кластеров не станет равным заданному значению. Размер кластера определяется как наибольшее расстояние между двумя входящими в него состояниями. В качестве расстояния может быть использовано взвешенное евклидово расстояние между средними гауссиан (для случая, когда состояния трифонов описываются одногауссовыми распределениями) либо другие виды метрик [6, 9].

2. Кластеризация на основе фонетического дерева. Этот метод [5, 10] использует бинарное фонетическое дерево решений, его преимущество — возможность моделирования трифонов, которых не было в обучающей выборке. Каждой вершине фонетического дерева соответствуют вопросы (с ответом „да/нет“), относящиеся к свойствам трифонов. Изначально все состояния, подлежащие кластеризации, образуют один класс (в вершине дерева). Далее на каждом шаге алгоритма, в зависимости от выбранного вопроса, происходит расщепление листа дерева на 2 класса. Вопрос в каждой вершине выбирается таким образом, чтобы максимизировать (локально) значение функции правдоподобия на обучающей выборке.

Предлагаемый алгоритм кластеризации состояний трифонов основан на стандартном агломеративном подходе с использованием приведенной ниже метрики, позволяющей в процессе связывания состояний явно учитывать количество данных (векторов в обучающей выборке), приходящихся на отдельные кластеры, и регулировать степень влияния распределения данных по кластерам на параметры связывания. В настоящей работе исследуется задача кластеризации данных для построения модели GMM-HMM. Акустическими единицами речи здесь являются контекстные трифоны. Каждый трифон состоит из трех состояний. Состояния трифонов описываются смесями гауссовых распределений.

1. Метрика должна отражать специфику данных рассматриваемой задачи. Сначала введем расстояние между двумя гауссианами. Расстояние от гауссианы G_1 до G_2 определим следующим образом с помощью расстояния Махаланобиса:

$$\rho(G_1, G_2) = \sqrt{\sum_{i=1}^{n} \frac{\left(\mu_{1i} - \mu_{2i}\right)^2}{\left(\sigma_{1i}\right)^2}}. \quad (1)$$

Здесь μ_{ji} — координата i среднего вектора гауссианы G_j, σ_{1i} — элемент диагональной ковариационной матрицы гауссианы G_1, n — размерность вектора признаков.

Для того чтобы расстояние было симметричным, используем следующую метрику:

$$\overline{\rho}(G_1, G_2) = \frac{1}{2}\left(\rho(G_1, G_2) + \rho(G_2, G_1)\right). \quad (2)$$

Процедура кластеризации состояний трифонов с использованием предложенной метрики включает следующие шаги. В начале процесса связывания все состояния трифонов описываются одногауссовыми распределениями. Далее происходит последовательное объединение гауссиан в классы, при этом на каждом шаге объединяются наиболее близкие классы.

Близость классов определим следующим образом: каждый класс C описывается множеством гауссиан его состояний, расстояние между двумя классами C_1 и C_2 вычисляется по формуле:

$$\rho(C_1, C_2) = \frac{1}{\sum\limits_{\substack{k \in C_1 \\ m \in C_2}} \left(N_k + N_m\right)} \sum_{\substack{k \in C_1 \\ m \in C_2}} \left(N_k + N_m\right) \overline{\rho}(G_k, G_m). \quad (3)$$

Здесь N_k — число векторов в обучающей выборке для состояния, которому соответствует гауссиана G_k.

2. Учет количества данных при кластеризации состояний трифонов. Для преодоления проблемы сбалансированности данных предлагается модифицировать метрику расстояний между классами (3) следующим образом:

$$\rho_{\text{bal}}(C_1, C_2) = \rho(C_1, C_2)(N_1 + N_2)^p. \quad (4)$$

Здесь p — степень влияния количества данных на метрику связывания, N_i — число векторов в кластере C_i.

Важно отметить, что оценки статистик (и расстояний) для состояний трифонов, у которых слишком мало данных в обучающей выборке, недостоверны. Поэтому в алгоритме кластеризации были выделены два этапа:

1) *слияние всех достаточно малых классов.* Если при выборе классов для объединения число векторов, приходящихся на какой-либо из них, меньше заданного порога thr, то расстояние между ними умножается на малое положительное число ε (например, $\varepsilon = 10^{-3}$);

2) *кластеризация всех остальных классов.*

Эти два этапа можно совместить, модифицировав метрику расстояний между классами следующим образом:

$$\rho^*(C_1, C_2) = \begin{cases} \rho_{\text{bal}}(C_1, C_2)\varepsilon, & \text{если } \min\left\{N_1, N_2\right\} < \text{thr}, \\ \rho_{\text{bal}}(C_1, C_2), & \text{если } \min\left\{N_1, N_2\right\} \geq \text{thr}. \end{cases} \quad (5)$$

Эксперименты. Цель экспериментов — установить, как влияет учет количества данных при связывании состояний на качество акустических моделей.

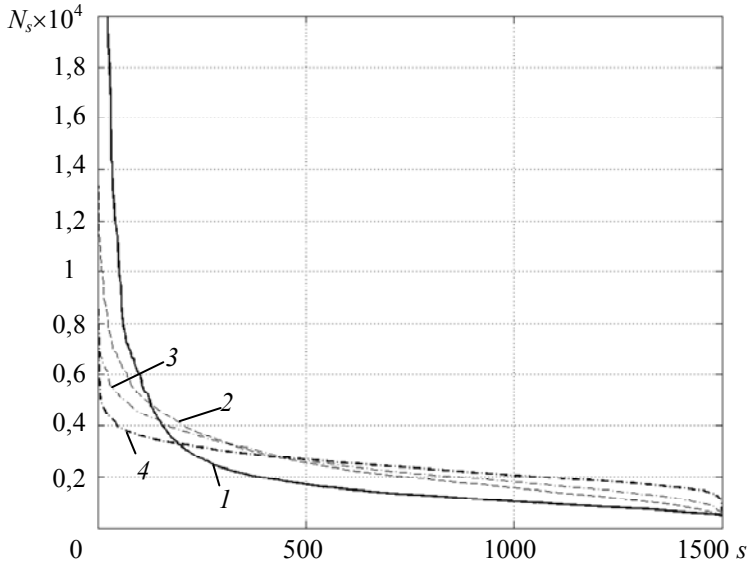
Речевая база данных для обучения состоит из фонограмм русской речи (как чтения, так и спонтанной) более чем 260 дикторов; тип канала — микрофон (16 кГц). Из исходной базы были сформированы четыре обучающие выборки размером 160, 80, 40 и 20 часов речи. Для каждой обучающей выборки были построены четыре акустические модели с разными значениями параметра p (см. (4)): 0; 0,4; 0,7; 2 (p=0 соответствует случаю, когда при кластеризации степень сбалансированности кластеров по количеству данных не учитывается).

Для построения акустических моделей использовались признаки MFCC+Δ+ΔΔ [11]: 12 мел-частотных кепстральных коэффициентов (MFCC, mel-frequency cepstral coefficients) и энергия вместе с производными первого и второго порядка от этих величин по времени. Фонетический алфавит состоит из 52 фонем русского языка и паузы. Во всех экспериментах количество связанных состояний трифонов равно 1500. Параметры обучения акустических моделей (за исключением p) во всех экспериментах совпадают, максимальное число гауссиан в состоянии 25. Тесты распознавания проводились на выборке из 1000 файлов общей длительностью 81 мин. Словарь для распознавания содержит 5400 словоформ. В качестве метрики для оценки качества акустических моделей используется пословная ошибка (Word Error Rate, WER), которая вычисляется следующим образом [12]:

$$\mathrm{WER} = \frac{S + D + I}{N} \cdot 100 .$$

Здесь S, D и I обозначают соответственно число ошибок замен, пропусков и вставок слов при распознавании; N — число слов в тексте, который был произнесен.

На рисунке для одной из обучающих выборок показано распределение количества векторов по состояниям трифонов после кластеризации для разных значений p (*1* — p=0; *2* — 0,4; *3* — 0,7; *4* — 2). На оси абсцисс приведены номера связанных состояний (s) после их упорядочивания в порядке убывания количества векторов N_s, приходящихся на эти состояния после кластеризации.



Результаты экспериментов приведены в таблице. Для каждой выборки показана разница (ΔWER) лучших результатов с учетом сбалансированности данных (p>0) и без ее учета (p=0), $\Delta\mathrm{WER}_{\mathrm{rel}}$ — относительные значения этой разницы. Для p=0,4 и 0,7 в среднем результаты получаются лучше, чем для p=0. Это подтверждает гипотезу о том, что возможность явного влияния на сбалансированность классов при кластеризации позволяет повысить качество аку-

стических моделей. Однако для случая, когда размер базы составляет всего 20 часов, улучшения при $p>0$ не наблюдается.

Результаты распознавания для разных значений p и различного объема обучающих выборок

Размер базы, ч	p	WER	ΔWER	ΔWER_{rel}
160	0	35,4	—	—
	0,4	34,2	—	—
	0,7	33,9	1,5	4,2
	2	34,4	—	—
80	0	35,2	—	—
	0,4	34,0	1,2	3,4
	0,7	34,1	—	—
	2	34,6	—	—
40	0	34,8	—	—
	0,4	33,8	1,0	2,9
	0,7	34,2	—	—
	2	34,4	—	—
20	0	33,8	—	—
	0,4	33,9	—	—
	0,7	33,8	0,0	0,0
	2	34,0	—	—

Заключение. В работе предложен метод учета количества данных при кластеризации состояний трифонов в процедуре обучения GMM-HMM для систем автоматического распознавания речи. Эксперименты показали, что предложенный метод позволяет повысить качество акустических моделей и при правильном выборе степени влияния количества данных позволяет уменьшить WER на 3—4 % относительно исходного значения. Уменьшения WER при учете количества данных удалось достичь только для обучающих выборок объемом не менее 40 часов. Используемая в статье базовая метрика (3) отличается от таких распространенных метрик, как расстояние Евклида, Бхатачария [9] и других [5, 6], но подход (4) можно обобщить на случай других метрик. При использовании базовой метрики для кластеризации состояний без взвешивания с учетом количества данных в классах будут неточными модели, для которых было мало данных в обучающей выборке. Введение метрики, приводящей к сбалансированности классов по количеству обучающих данных, позволяет построить более надежные модели для тех состояний, которые в первом случае оказались плохо обученными, но в то же время может сделать менее точным разделение между остальными моделями. Для лучшей кластеризации необходим выбор параметров метода, обеспечивающий наилучшее соотношение надежности моделей и точности их разделения.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Irtza S., Hussain S.* Minimally balanced corpus for speech recognition // Proc. 1st IEEE Intern. Conf. on Communications, Signal Processing, and their Applications (ICCSPA). 2013. P. 1—6.

2. *Guo X., Yin Y., Dong C., Yang G., Zhou G.* On the class imbalance problem // Proc. IEEE 4th Intern. Conf. on Natural Computation (ICNC'08). 2008. Vol. 4. P. 192—201.

3. *Garcia-Moral A. I., Solera-Ureña R., Peláez-Moreno C., Díaz-de-María F.* Data balancing for efficient training of hybrid ANN/HMM automatic speech recognition systems // IEEE Transact. on Audio, Speech, and Language Processing. 2011. Vol. 19, N 3. P. 468—481.

4. *Darjaa S., Cernak M., Trnka M., Rusko M., Sabo R.* Effective Triphone Mapping for Acoustic Modeling in Speech Recognition // Proc. INTERSPEECH. 2011. P. 1717—1720.

5. *Young S. J., Odell J. J., Woodland P. C.* Tree-based state tying for high accuracy acoustic modelling // Proc. of the Workshop on Human Language Technology. Association for Computational Linguistics. 1994. P. 307—312.

6. The HTK book / *Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Woodland P.* // Cambridge University Engineering Department. 2002.

7. *Aubert X., Beyerlein P., Ullrich M.* A bottom-up approach for handling unseen triphones in large vocabulary continuous speech recognition // Proc. IEEE 4th Intern. Conf. on Spoken Language (ICSLP 96). 1996. Vol. 1. P. 14—17.

8. *Park J., Ko H.* Effective acoustic model clustering via decision-tree with supervised learning // Speech communication. 2005. Vol. 46. N 1. P. 1—13.

9. *Mak B., Barnard E.* Phone clustering using the Bhattacharyya distance // Proc. IEEE 4th Intern. Conf. on Spoken Language (ICSLP 96). 1996. Vol. 4. P. 2005—2008.

10. *Wang G., Sim K. C.* An investigation of tied-mixture GMM based triphone state clustering // Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). 2012. P. 4717—4720.

11. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение, 2013. Т. 56, № 2. С. 47—51.

12. *Khokhlov Y., Tomashenko N.* Speech Recognition Performance Evaluation for LVCSR System // Proc. 14th Intern. Conf. “SPEECH and COMPUTER” (SPECOM 2011). Kazan. Russia. 2011. P. 129—135.

Сведения об авторах

Наталья Александровна Томашенко — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ“, Санкт-Петербург; младший научный сотрудник;
E-mail: tomashenko-n@speechpro.com

Юрий Юрьевич Хохлов — ООО „ЦРТ“, Санкт-Петербург; ведущий программист;
E-mail: khokhlov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 004.934

Г. А. Черных, М. Л. Кореневский, К. Е. Левин,
И. А. Пономарева, Н. А. Томашенко

КРОССВАЛИДАЦИОННЫЙ КОНТРОЛЬ СОСТОЯНИЙ ПРИ ОБУЧЕНИИ АКУСТИЧЕСКИХ МОДЕЛЕЙ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

Предложен метод, позволяющий при обучении скрытых марковских моделей, входящих в состав систем автоматического распознавания речи, оптимизировать число компонент в гауссовых смесях состояний. Применение метода повышает качество и скорость распознавания речи системой.

Ключевые слова: *автоматическое распознавание речи, скрытые марковские модели, кроссвалидационный контроль, гауссова смесь.*

Введение. Скрытые марковские модели (hidden Markov models, HMM) являются важной составляющей многих систем автоматического распознавания речи [1]. При обучении HMM-моделей часто возникает проблема недостаточного количества обучающих данных. В подобных случаях необходимо соблюдать баланс между сложностью моделей и их обобщающей способностью. Чрезмерно сложные модели, содержащие множество гауссиан в гауссовых смесях состояний, склонны к переобучению и, как следствие, теряют свою эффективность на данных, не входящих в обучающую выборку.

Предлагаемый в настоящей работе подход позволяет оптимизировать сложность моделей в соответствии с речевыми данными, в результате чего снижается вычислительная сложность алгоритма и повышается качество распознавания речи. Этот подход был успешно применен при обучении акустических моделей системы автоматического распознавания речи ООО „ЦРТ“ для русского языка, основой которой является тандемная структура [2], где сформированные нейронными сетями и подвергнутые дополнительной обработке [3] акустические признаки речи используются для обучения HMM-моделей.

Кроссвалидационный контроль используется во многих задачах машинного обучения с целью исключения эффекта переобучения, например, при обучении искусственных нейронных сетей в ходе подстройки параметров. Однако обучение HMM-модели сопровождается усложнением самого субъекта (модели) в результате последовательного расщепления гауссовых смесей. Кроссвалидационный контроль можно применять и при обучении HMM-моделей, но о его использовании в таких задачах практически не упоминается в литературе; в качестве критерия останова процесса обучения традиционно предлагается проводить тестирование качества распознавания на небольшой выборке [4]. Представленный в настоящей работе метод позволяет эффективно решать задачу построения гауссовых смесей и своевременно останавливать обучение в целом.

Метод основан на использовании специальных кроссвалидационных критериев, регулирующих количество компонент в гауссовых смесях состояний при итеративном обучении моделей. Критерии вводятся для каждого состояния в отдельности и строятся по отдельной части речевой базы, не принимающей участия в обучении моделей. Обучение HMM с увеличением количества гауссиан в состояниях заключается в чередовании эпох* „спокойного“ обучения, когда происходит только подстройка параметров без усложнения моделей, и одномоментных расщеплений гауссиан состояний [3]. Решение о целесообразности расщепления

* Эпоха (*англ.* epoch) — однократный проход по всему обучающему множеству.

каждого из состояний принимается на основе динамики кроссвалидационных критериев в конце эпохи обучения (непосредственно перед моментом очередного расщепления). Таким образом, количество гауссиан увеличивается не для всех, а только для отобранных состояний. Кроме того, в ходе обучения возможно и ухудшение динамики критерия в некоторых состояниях, например, вследствие эффекта переобучения, вызванного недостатком данных. В этом случае происходит возврат состояний в их предыдущие реализации, для которых соответствующие кроссвалидационные критерии показывали бóльшую эффективность обучения.

Стандартная процедура обучения HMM-модели состоит из следующих основных этапов (см., например, [4]):

1) обучение монофонных одногауссовых моделей;

2) получение несвязанных одногауссовых трифонных моделей посредством клонирования монофонных моделей;

3) дообучение несвязанных одногауссовых трифонных моделей;

4) связывание состояний одногауссовых трифонных моделей с использованием дерева решений;

5) обучение трифонных моделей со связанными состояниями с последовательными расщеплениями состояний и получение многогауссовых трифонных моделей [5].

Предлагаемый в настоящей работе метод может применяться на пятом этапе обучения, во время которого происходит последовательное наращивание количества гауссиан в связанных состояниях трифонных моделей.

Прежде чем перейти к описанию алгоритма, отметим две особенности предлагаемого подхода. Во-первых, сбор статистики для обновления акустических моделей производится с помощью разновидности алгоритма Баума—Уэлша, так называемого алгоритма „точного совпадения“ (exact-match), используемого при дискриминативном обучении [6]. В отличие от канонического алгоритма Баума—Уэлша, требующего только информацию о последовательностях фонем, версии „точного совпадения“ необходима пофонемная разметка обучающих данных [7]. При этом алгоритм „прямого-обратного“ хода (forward-backward) [1], который является основой метода сбора статистики для обновления HMM-моделей, применяется не к каждой фразе целиком, а к интервалам, соответствующим положению отдельных фонем этой фразы в разметке (при дискриминативном обучении эти интервалы соответствуют ребрам фонемной сети гипотез). Знание относительных вероятностей посещения (occupancy probability) [1] на интервалах, их временных границ и последовательности фонем позволяет получить абсолютные значения вероятностей, используемых для обновления моделей. Подобный подход значительно ускоряет процедуру обучения и в общем случае улучшает качество моделей, так как в процессе разметки из обучающего множества могут быть исключены заведомо ошибочные данные. Во-вторых, поскольку алгоритм призван оптимизировать количество гауссиан, добавляется только по одной гауссиане в состояние при его расщеплении.

Кроссвалидационный контроль и его критерии. При обучении моделей речевые данные разделяются на две части: первая (обучающая) применяется непосредственно для обучения, вторая (кроссвалидационная) используется только для вычисления критериев. Процедура вычисления кроссвалидационного критерия для состояния состоит в следующем. К кроссвалидационным данным, аналогично тому как это делается на обучающих данных с учетом специфики подхода „точного совпадения“, применяется алгоритм „прямого-обратного“ хода. Для каждого состояния s HMM-модели, присутствующего в последовательности фонем обрабатываемой фразы, вычисляются вероятности его наблюдения в каждый из моментов времени, соответствующих фонеме. Набор вероятностей наблюдения образует кусочную функцию дискретного времени $L_i^s(t)$, где i — номер фразы. Из этих значений вероятности выбираются все локальные максимумы $\hat{t}_1^s,\ldots,\hat{t}_{M(s,i)}^s$ и помещаются в „аккумулятор“

состояния. После обработки всех фраз, входящих в кроссвалидационную выборку, вычисляется среднее по „аккумулятору“ состояния, что и является искомым критерием:

$$C(s) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{M(s,i)} L_i^s\left(\hat{t}_k^s\right)}{\sum_{i=1}^{N} M(s,i)}, \quad (1)$$

где N — число фраз в кроссвалидационной выборке.

Вместе с индивидуальными кроссвалидационными критериями состояний, позволяющими оптимизировать количество гауссиан в состояниях, вводится общий кроссвалидационный критерий, который используется для фиксации времени останова обучения. Критерий вычисляется усреднением значений правдоподобия, определяемых с помощью алгоритма „прямого-обратного“ хода по полным фразам кроссвалидационной выборки. Пусть T_i — число кадров в i-й фразе кроссвалидационной выборки и LLH_i — суммарная вероятность, тогда критерий:

$$C = \frac{1}{N} \sum_i \frac{\mathrm{LLH}_i}{T_i}. \quad (2)$$

Поскольку обучение происходит по критерию максимального правдоподобия, то значения аналогичных критериев, вычисляемые на основе обучающей выборки, в процессе обучения должны возрастать. Однако динамика критериев, рассчитанных по кроссвалидационной выборке, в ходе обучения и наращивания количества гауссиан может быть довольно сложной и необязательно монотонной. Описываемый далее алгоритм расщепления гауссовых смесей эффективно обеспечивает рост критериев в целом.

Стратегия расщепления гауссовых смесей. Предлагаемый алгоритм расщепления характеризуется следующими основными особенностями:

— слежение за каждым состоянием в отдельности;

— использование резервных копий состояний для возврата к более удачным реализациям;

— единовременные расщепления и возвраты состояний с последующими эпохами „спокойного“ обучения.

Алгоритм функционирует следующим образом. После нескольких первичных итераций обучения одногауссовых моделей создаются резервные копии всех состояний, и для каждого состояния запоминается текущее значение его кроссвалидационного критерия (1). Далее производится расщепление всех состояний и выполняется несколько итераций обучения. В конце эпохи обучения сравниваются текущие значения кроссвалидационных критериев состояний с их предыдущими записанными значениями. Если после эпохи обучения кроссвалидационный критерий состояния уменьшился, то это состояние возвращается назад к его резервной реализации. Для всех остальных состояний с возросшим кроссвалидационным критерием применяются те же операции, что и перед первым расщеплением: резервное копирование текущих реализаций состояний и запись текущих значений кроссвалидационных критериев, при этом предыдущие резервные копии удаляются. Таким образом, посредством использования резервных копий состояний обеспечивается рост кроссвалидационных критериев.

По мере роста числа гауссиан количество состояний, возвращаемых к их предыдущим реализациям, начинает превалировать над числом состояний, для которых необходимо увеличить количество гауссиан в смеси. В конце обучения, когда почти все состояния имеют оптимальное количество гауссиан, может возникнуть ситуация, при которой после последнего расщепления рост критериев сменится спадом, поэтому обучение необходимо останавливать непосредственно после откатов состояний к их резервным копиям. Останов обучения производится либо после достижения желаемого количества расщеплений, которое должно быть не

меньше предполагаемого числа гауссиан в состояниях, либо когда прирост ΔC критерия (2) в ходе обучения перестанет превышать заданное значение.

Производить расщепление или возврат к резервной копии каждого состояния в отдельности вместо единовременной процедуры расщеплений и откатов нецелесообразно, поскольку постоянные скачкообразные изменения состояний крайне замедляют процедуру обучения в целом.

Результаты. Эксперименты проводились на базе русской речи SpeechDat(E) [8], содержащей телефонные записи фонетически сбалансированных предложений, слов, словосочетаний, чисел и числовых последовательностей. Процедура сравнения эффективности традиционного обучения с подходом, предложенным в настоящей работе, разбита на две группы экспериментов: на полной базе (около 67 часов речи) и 15 %-ной случайной выборке с целью моделирования недостатка данных (примерно 10 часов речи), который в случае применения метода обучения без контроля состояний быстро приводит к эффекту переобучения и значительно ухудшает качество распознавания посредством обученных подобным образом акустических моделей. Связывание состояний трифонных моделей проводилось с помощью дерева решений [1]. Всего имелось 18 800 связанных состояний для моделей, обученных по полной базе, и около 10 000 связанных состояний для моделей, обученных по 15 %-ной выборке.

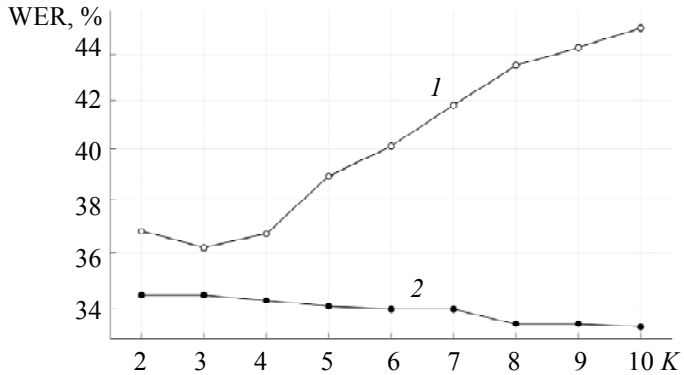
Тестирование качества распознавания речи велось по словарю в 12 500 слов без использования языковой модели. Параметры системы распознавания подбирались таким образом, чтобы обеспечить одинаковую скорость распознавания акустическими моделями, обученными обоими сравниваемыми методами. Результаты обучения (пословная ошибка WER и среднее число гуссиан $\overline{N}_g$) на полном объеме данных на 15 %-ной выборке приведены в таблице.

Результаты обучения

Метод обучения	Полная база		15 %-ная выборка	
	WER, %	$\overline{N}_g$	WER, %	$\overline{N}_g$
Традиционный	34,3	11	45,2	11
Предлагаемый	32,2	3,2	33,4	2,1

Из таблицы видно, что по сравнению с методом обучения без кроссвалидационного контроля предложенный метод дает ощутимое преимущество, более отчетливо проявляющееся в случае недостаточного количества данных для обучения при значительно меньшем количестве гауссиан в состояниях.

На рисунке приведены зависимости ошибки обучения на 15 %-ной выборке от числа K произведенных расщеплений гауссовых смесей. Отчетливо видно, что без кроссвалидационного контроля состояний обычный метод обучения (*1*) приводит к ухудшению качества распознавания, что является свидетельством эффекта переобучения, которого не наблюдается при предложенном подходе (*2*).



Отметим, что индивидуальный контроль состояний позволил достичь приемлемого качества обучения HMM-модели даже на небольшой части обучающей базы. Результаты, полученные этими моделями, уступают полученным на полной базе, только 1 % (абсолютных).

Заключение. В настоящей работе предложена методика обучения акустических моделей с последовательным наращиванием количества гауссиан в состояниях HMM-модели, контролируемым на уровне состояний посредством введения критериев, вычисляемых на кроссвалидационной выборке речевой базы данных. Полученные результаты демонстрируют эффективность метода с точки зрения повышения качества обученных HMM одовременно с сокращением сложности самих моделей вследствие оптимального выбора количества гауссиан.

В качестве развития метода следует рассмотреть возможность использования более тонких кроссвалидационных критериев, поскольку очевидно, что предложенный метод их вычисления — это лишь один из возможных способов ввести меру, которая бы позволила судить о необходимости дальнейших расщеплений состояний. В частности, применение вероятностей посещения, возвращаемых алгоритмом „прямого-обратного“ хода, по-видимому, может повысить эффективность метода. Процедура применения резервных копий состояний для возвратов в случае необходимости к их предыдущим реализациям также может быть модифицирована. Использование нескольких резервных копий состояний в совокупности с алгоритмами отбора этих копий также может повысить качество обучения.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Young S., Evermann G., Hain, T., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland P.* The HTK Book. Cambridge University Engineering Dept. [Электронный ресурс]: <http://htk.eng.cam.ac.uk/docs/docs.shtml>.
2. *Schwarz P.* Phoneme recognition based on long temporal context. PhD thesis. – Brno University of Technology, 2008 [Электронный ресурс]: <http://www.fit.vutbr.cz/~schwarzp/publi/thesis.pdf>.
3. *Andrew J. N.* Model Reduction via the Karhunen-Loeve Expansion. Part I: An Exposition. Technical Report. University of Maryland [Электронный ресурс]: <http://drum.lib.umd.edu/bitstream/1903/5751/1/TR_96-32.pdf>.
4. *Huang X., Acero A., Hon H.W.* Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, 2001. 1008 p.
5. *Khokhlov Y., Kiselev V., Tampel I., Tatarnikova M.* Phone Recognition Driven Method for Creating Context-Dependent Phones // Proc. of the 12th Intern. Conf. on Speech and Computer (SPECOM 2007). Moscow, 2007. P. 499—502.
6. *Povey D.* Discriminative training for large vocabulary speech recognition. PhD thesis. – Cambridge University Engineering Dept, 2003 [Электронный ресурс]: <https://sites.google.com/site/dpovey/phd_2003.pdf?attredirects=0>.
7. *Khokhlov Y, Tomashenko N.* Speech Recognition Performance Evaluation for LVCSR System // Proc. of the 14th Intern. Conf. on Speech and Computer (SPECOM 2011). Kazan, 2011. P. 129—135.
8. *Heuvel V. H., Boudy J., Bakcsi Z., Cernocky J., Galunov V., Kochanina J., Majewski W., Pollak, P, Rusko M., Sadowski J., Staroniewicz P., Tropf H.S.* SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed // Proc. of 7th Europ. Conf. on Speech Communication and Technology (Interspeech 2001). Scandinavia, 2001. P. 2059—2062.

Сведения об авторах

Герман Анатольевич Черных — канд. физ.-мат. наук; ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; Санкт-Петербургский государственный университет; кафедра проблем конвергенции естественных и гуманитарных наук; доцент; E-mail: chernykh@speechpro.com

Максим Львович Кореневский — канд. физ.-мат. наук; ООО „ЦРТ-Инновации“, Санкт-Петербург; научный сотрудник; E-mail: korenevsky@speechpro.com

Кирилл Евгеньевич Левин — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела распознавания речи; E-mail: levin@speechpro.com

Ирина Александровна Пономарева — ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; E-mail: ponomareva@speechpro.com

Наталья Александровна Томашенко — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ“, Санкт-Петербург; младший научный сотрудник; E-mail: tomashenko-n@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 81'322.6

П. Г. Чистиков, О. Г. Хомицевич, С. В. Рыбин

СТАТИСТИЧЕСКИЕ МЕТОДЫ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ МЕСТ И ДЛИТЕЛЬНОСТИ ПАУЗ В СИСТЕМАХ СИНТЕЗА РЕЧИ

Рассмотрены статистические методы определения местоположения и длительности пауз в системе синтеза речи. Применение таких методов позволяет добиться лучших результатов по сравнению с использованием алгоритмов, основанных на правилах.

Ключевые слова: *пауза, синтез речи, статистические модели.*

Введение. Корректная просодическая разметка в системах синтеза речи необходима для естественного звучания синтезированной речи. Обычно достаточно длинные предложения разбиваются на отдельные фрагменты, которые разделяются паузами. Такие паузы делают речь более понятной и естественной, разрешая неоднозначные трактовки смысла предложений.

Многие системы синтеза речи при определении мест пауз опираются только на знаки препинания. Однако большие участки текста, расположенные между этими знаками, могут звучать монотонно и осложнять восприятие речи, что делает актуальной задачу определения мест пауз на подобных участках. При синтезе русской речи дополнительно возникает другая проблема — знаки пунктуации традиционно используются для обособления различных вводных конструкций, таких как „может быть“, „конечно“ и т.д., которые не выделяются паузами в устной речи.

Кроме того, системы синтеза речи должны не только определять места пауз, но и их продолжительность как внутри предложений, так и между ними. Самым простым решением этой задачи является задание различных констант, регламентирующих длительность пауз. Но так как длительность естественных (в речи человека) пауз является очень вариативной величиной, необходим специальный метод, позволяющий вычислять длительность пауз в зависимости от контекста и структуры предложения [1].

Использование пауз в естественной речи зависит от ряда факторов. Наиболее значимым из них является синтаксическая структура предложения: паузы зачастую располагаются между синтаксически связными компонентами [2, 3]. Однако длина предложения, семантика определенных слов и другие особенности также имеют значение [4]. В системах синтеза речи эти факторы могут быть учтены путем задания правил, определяющих, после какого слова в предложении должна стоять пауза [5, 6], или путем обучения статистических моделей на большом речевом корпусе, на основе которых будут вычисляться вероятности наличия пауз после того или иного слова [7, 8].

В системе синтеза русской речи компании ООО „ЦРТ“ для определения мест пауз применяются алгоритмы, основанные на правилах, а длительность пауз определяется на основе определенных констант [9]. Такой подход достаточно хорош, однако невозможно учесть все случаи, встречающиеся в различных текстах, помимо того, разработка подобных правил для новых (для системы) языков требует большого количества времени. Преимущество методов машинного обучения — простота применения, при наличии размеченного речевого корпуса достаточного объема. Ожидается, что статистические модели будут более детально эмитировать поведение человека, нежели правила, основанные на знаниях экспертов. В настоящей работе исследуются пути, позволяющие улучшить алгоритм, основанный на правилах, путем использования статистических методов и просодического анализа.

Классификаторы CART и RF. Классификатор CART [10] применяется для определения местоположения и длины пауз: определяется длительность паузы между словами (там, где она равна нулю или меньше заданного порога, пауза отсутствует). Также этот классификатор использовался только для определения длины пауз: в этом случае предсказывается длительность паузы только между теми словами, где она была поставлена на предыдущих этапах обработки текста. Классификатор RF [11] применялся только для определения местоположения пауз.

CART — рекурсивный метод разбиения набора данных на основе минимизации функции:

$$G(C_1, C_2) = \frac{D(C_1)T(C_1) + D(C_2)T(C_2)}{T(C_1) + T(C_2)}, \qquad (1)$$

где

$$D(C) = \frac{\sum_{i=1}^{|C|}\sum_{j=i}^{|C|} d(U_i, V_j)}{T(C)},$$

$$T(C) = \frac{1}{2}|C|\left(|C|-1\right).$$

$|C|$ — размер кластера C, $d(U,V)$ — расстояние между векторами признаков U и V, критерием останова служит минимальное число элементов в кластере (в настоящей работе — три).

RF выполняет классификацию данных на основе множества признаков путем создания иерархии („деревьев“) запросов на основе предсказанных значений признаков в каждой точке. Лист каждого дерева содержит информацию обо всех наблюдениях характеризуемой величины, признаки которой лежат в одной области значений. В разработанной экспериментальной системе применяется „лес решений“, содержащий 100 деревьев, где каждое дерево построено на 60 % случайно выбранных данных, это снижает чувствительность алгоритма к шуму в обучающих данных.

Описание эксперимента. Использовался размеченный звуковой корпус, состоящий из записей девяти дикторов (4 мужчин и 5 женщин). Записанный материал — русскоязычная художественная литература и новостные статьи. Итоговая база данных содержит более 50 часов речи, включая 38 000 пауз, для выполнения тестов она была разделена на тестовую и обучающую выборки в отношении 85 к 15 % соответственно.

При обучении модели определения местоположения учитывались только паузы внутри предложений исходя из того, что границы предложений известны (текст с высокой точностью разделяется на предложения процедурой нормализации текста, встроенной в систему синтеза речи). Для определения длительности модель обучалась на основе данных о паузах как внутри предложений, так и между ними.

Для решения задачи классификации использовались

— пунктуация: знак препинания после текущего слова, после двух предыдущих и после двух последующих слов;

— число слов и слогов: слов и слогов в предложении, слов и слогов от предыдущей паузы до текущего слова и от текущего слова до конца предложения;

— грамматические признаки: часть речи, падеж, признак „является ли слово собственным существительным“ (имена, названия и т.д.) — данная информация получается с помощью морфологического словаря, входящего в состав системы синтеза речи;

— признаки согласования: согласуется ли грамматическая форма текущего слова с формой двух последующих слов;

— регистр первой буквы слова: является ли она в двух предыдущих, в текущем или двух последующих словах заглавной или нет.

Как при обучении, так и при тестировании, для снижения числа ошибок вычисления грамматических признаков необходима процедура разрешения неоднозначности для слов-омонимов и омографов (замОк — зАмок). В настоящей статье используется предложенный в работе [12] подход, точность работы которого составляет 96 %.

Определение местоположения пауз. В табл. 1 сравниваются результаты использования метода автоматического определения мест пауз (CART и RF) с результатами работы базового, основанного на правилах, подхода, встроенного в систему синтеза русской речи, использованную при выполнении настоящей работы. Тестовая выборка содержит 47 819 пар слов внутри предложения, между которыми возможна пауза, и 6186 пауз (аналогичные показатели для обучающей выборки равняются 264 336 и 32 630 соответственно).

Таблица 1

Результаты определения местоположения пауз

№	Параметр	Базовый алгоритм	CART	RF
1	Правильных границ между словами	43 254 (90 %)	44 358 (93 %)	44 865 (94 %)
2	Правильных пауз	5042 (82 %)	5176 (84 %)	4695 (76 %)
3	Ложных срабатываний	3421 (55 %)	2451 (40 %)	1463 (24 %)
4	Ложных отклонений	1144 (18 %)	1010 (16 %)	1491 (24 %)
5	Эффективность, %	82	84	76
6	Точность, %	60	68	76
7	F-мера, %	69	75	76

В табл. 1 в строке № 1 указано число корректно идентифицированных „пауз“ или „не пауз“ среди всех пар слов тестовой выборки; в строке № 2 указаны корректно определенные паузы. Эта мера использовалась в работе [7] и приведена здесь для сравнения, в равной степени как и показатели ложных срабатываний (FA), ложных отклонений (FR) и F-меры (F-score). Результаты работы обоих классификаторов превосходят результаты работы базовой версии алгоритма: значение F-меры больше, а показатели ложных срабатываний и ложных отклонений более сбалансированы. Результаты работы CART и RF различаются незначительно, однако, что немаловажно, уровень FA/FR при применении RF может быть настроен произвольным образом.

Результаты работы сравнимы с результатами, описанными в литературе. Так, например, для английского языка количество „правильных границ между словами“ составляет 91,1 %, а F-мера равна 71,9 % [7]. В работе [8] данный показатель составляет 74,4 %.

Определение длительности пауз. При проведении экспериментов по определению длительности пауз классификатор изначально был обучен для работы со всеми типами пауз. Затем было решено определять длительность пауз между предложениями и внутри них отдельно. Для представления результатов работы системы использовалась мера NRMSD (нормализованное среднее квадратичное отклонение, Normalized Root-Mean-Square Deviation), схожая с мерой, применяемой в работе [1]; чем это значение ближе к нулю, тем результат лучше.

В табл. 2 приведены результаты применения обобщенной (создана для определения всех пауз в наборе данных) и специализированных моделей (две различные модели для определения пауз внутри и между предложениями).

Таблица 2

Результаты определения длительности пауз

Модель	Паузы внутри предложения	Паузы между предложениями
Обобщенная	0,25	0,23
Специализированные	0,19	0,16

Проанализировав таблицу, можно сделать вывод, что специализированные модели позволяют более точно описать параметры пауз как между предложениями, так и внутри них.

Интеграция с системой синтеза речи. Как было сказано выше, базовая версия алгоритма определения местоположения пауз основана на правилах, а длительность пауз задается соответствующими константами. Паузы делятся на четыре типа в зависимости от их длины: один тип пауз между предложениями и три — внутри предложений. Такой подход является достаточно статичным для ритмики предложений. Как следствие, было решено использовать алгоритм, основанный на статистических методах, в первую очередь были внедрены модели, на основе которых отдельно определялась длительность пауз внутри предложений и между ними. Затем был интегрирован статистический подход для определения мест пауз. В итоге была получена система, работа которой включает следующие этапы:

1) расстановка пауз согласно знакам препинания (за исключением мест, где знаки препинания не предполагают пауз);

2) разбиение паузами длинных цепочек слов без знаков препинания на основе статистической модели (CART или RF);

3) определение длительности полученных пауз на основе статистики.

Сравнив подходы с использованием классификаторов CART и RF, можно отметить следующее. Очевидным преимуществом использования CART является маленький размер модели, что является важным показателем при реализации системы синтеза речи. Однако, как показано в табл. 1, RF дает лучшие результаты при определении местоположения пауз. Более того, не все ошибки одинаково критичны: в некоторых случаях пауза недопустима, в то время как в других имеет право быть. В работе CART возникают более серьезные ошибки по сравнению с RF, хотя это может быть выявлено только на основе экспертных оценок. В основном ошибки CART заключаются в паузах внутри синтаксически связанных цепочек: после предлогов, союзов и других служебных слов, использующихся для связи последовательности слов; между модификатором (прилагательное, наречие и т.д.) и существительным или глаголом, к которому он относится. Такого рода ошибки практически отсутствуют при использовании классификатора RF. Помимо того, модель с использованием RF является более гибкой, поскольку она может быть настроена с целью увеличения или уменьшения количества пауз в синтезируемой речи, что может быть полезно для практических приложений системы синтеза речи. Например, увеличение количества пауз снижает темп речи.

Сравнение результатов работы полученной экспериментальной системы с базовой — сложная задача, она требует применения MOS-оценки (Mean Opinion Score), что является целью будущих исследований. Однако для определения предпочтений слушателей был проведен тест. Было выбрано 25 предложений, содержащих большие последовательности слов, не разделенные знаками препинания. Результаты оценивались 18 русскоязычными экспертами: 10 предпочли предложенную систему, 4 не смогли определить, что лучше, 4 предпочли базовую.

Заключение. В работе представлен подход к определению местоположения и длительности пауз в системе синтеза речи на основе статистических моделей. Результаты экспериментов показали, что модели, построенные на основе алгоритмов классификации данных CART и RF, в сравнении с основанным на правилах подходом дают лучшие результаты.

Модель CART допускает более критические, по сравнению с RF ошибки, в основном заключающиеся в лишних паузах между синтаксически связанными участками текста. Таким образом, в системе синтеза речи предпочтительно использовать модель RF. Такая экспериментальная система получила положительную оценку слушателей-экспертов.

Был проведен ряд экспериментов по определению длительности пауз на основе алгоритма CART: обнаружено, что он работает лучше в случае применения различных моделей для предсказаний длительности пауз внутри предложений и между ними. На основе экспертных оценок был сделан вывод, что это решение позволяет повысить естественность синтезированной речи.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Parlikar A., Black A. W.* Modeling Pause-Duration for Style-Specific Speech Synthesis // Proc. of Interspeech. Portland, OR, USA, 2012. P. 446—449.
2. *Bachenko J., Fitzpatrick E.* A computational grammar of discourse-neutral prosodic phrasing in English // Computational linguistics. 1990. Vol. 16 (3). P. 155—170.
3. *Tepperman J., Nava E.* Where should pitch accents and phrase breaks go? A syntax tree transducer solution // Proc. of Interspeech. Florence, Italy, 2011. P. 1353—1356.
4. *Zellner B.* Pauses and the temporal structure of speech // Fundamentals of speech synthesis and speech recognition / Ed. by *E. Keller*. Chichester: John Wiley, 1994. P. 41—62.
5. *Abney S.* Parsing by chunks // Principle-Based Parsing Computation and Psycholinguistics. 1991. Vol. 44. P. 257—278.
6. *Atterer M.* Assigning Prosodic Structure for Speech Synthesis: A Rule-based Approach // Proc. of Speech Prosody. Aix-en-Provence, 2002. P. 147—150.
7. *Black A. W., Taylor P.* Assigning phrase breaks from part-of-speech sequences // Computer Speech & Language. 1998. Vol. 12, N 2. P. 99—117.
8. *Busser B., Daelemans W., Bosch A. V. D.* Predicting phrase breaks with memory-based learning // Proc. of 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis. 2001. P. 29—34.
9. *Хомицевич О. Г., Соломенник М. В.* Автоматическая расстановка пауз в системе синтеза русской речи по тексту // Компьютерная лингвистика и интеллектуальные технологии: Матер. Междунар. конф. „Диалог“. М.: Изд-во РГГУ, 2010. Вып. 9 (16). С. 531—537.
10. *Loh W.-Y.* Classification and Regression Tree Methods // Encyclopedia of Statistics in Quality and Reliability. Wiley, 2008. P. 315—323.
11. *Breiman L. Cutler A.* Random Forests [Электронный ресурс]: <http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>.
12. *Хомицевич О. Г., Рыбин С. В., Аничкин И. М.* Использование лингвистического анализа для нормализации текста и снятия омонимии в системе синтеза русской речи // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 42—46.

Сведения об авторах

Павел Геннадьевич Чистиков — ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; E-mail: chistikov@speechpro.com

Ольга Гурьевна Хомицевич — PhD; ООО „ЦРТ“, Санкт-Петербург; старший научный сотрудник; E-mail: khomitsevich@speechpro.com

Сергей Витальевич Рыбин — канд. физ.-мат. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: rybin@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

СИСТЕМЫ ОБРАБОТКИ РЕЧЕВЫХ И АКУСТИЧЕСКИХ СИГНАЛОВ

УДК 621.391.037.372

С. В. Алейник, М. Б. Столбов

ОЦЕНКА ВРЕМЕННОГО СДВИГА МЕЖДУ АУДИОСИГНАЛАМИ С ИСПОЛЬЗОВАНИЕМ ИХ ОГИБАЮЩИХ

Предложен метод оценки временного сдвига между акустическими сигналами, записанными в условиях реверберации и нелинейных искажений, базирующийся на оценке кросскорреляции временных огибающих сигналов, проведено его сравнение с другими методами оценки временного сдвига.

Ключевые слова: *временной сдвиг, временная огибающая, кросскорреляция, речевой сигнал.*

Введение. Оценка временного сдвига (ВС) между двумя сигналами (обычно называемыми „основной“ и „опорный“) важна для решения многих задач обработки аудиосигналов [1—5]: например, при оценке направления прихода сигналов, учете задержки в алгоритмах двухканальной фильтрации и др.

Большинство способов определения ВС базируется на оценке меры „близости“ сигналов друг к другу: функции кросскорреляции (ФКК) сигналов, обобщенной кросскорреляции (generalized cross-correlation, GCC), евклидова расстояния между сигналами, а также методе преобразования фазы ФКК (phase transform, PHAT) и т.п. [6—8]. Ряд факторов, таких как реверберация, увеличение расстояния между приемниками аудиосигналов, нелинейные искажения сигналов, уменьшает сходство между сигналами, что приводит к снижению стабильности оценок ВС. На рис. 1 приведены оценки ФКК (R_x) аудиосигналов, записанных в помещении при расстоянии между основным и опорным микрофонами 1 метр (кривая *1*), 2 (*2*) и 3 (*3*). Видно, что с увеличением расстояния максимум ФКК сигналов существенно снижается.

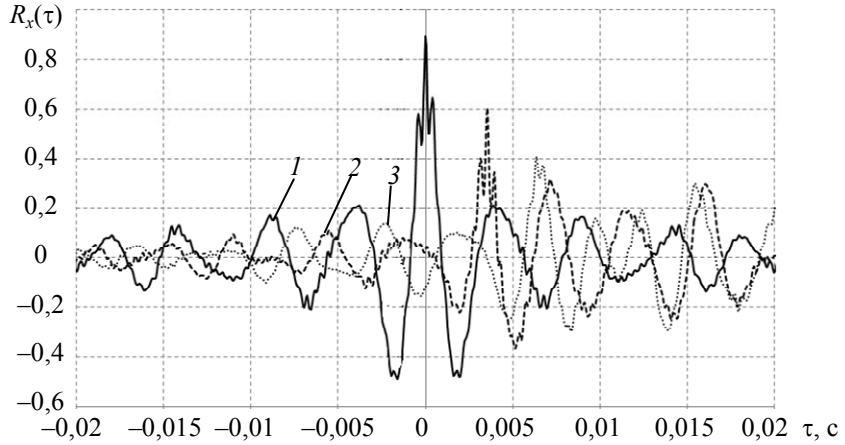


Рис. 1

В работе [9] показано, что оценка ВС на основе функции кросскорреляции временных огибающих сигналов (ФКО) дает хорошие результаты в случае сильных реверберационных искажений, назовем это методом корреляции огибающих (МКО). Обычно оценки ВС с использованием огибающих применяются в обработке коротких импульсных узкополосных сигналов в радиолокации и гидролокации [4, 7, 8, 10, 11], однако не для широкополосных аудиосигналов. Целью предлагаемой работы является описание алгоритма МКО, определение границ его применимости и оптимальных параметров.

Описание алгоритма. Оценка ВС в предлагаемом методе производится так же, как в методе ФКК. Однако сама ФКК вычисляется не по исходным сигналам, а по их временным огибающим, т.е.

$$\hat{\tau} = \arg\max(R_{a_1,a_2}(\tau)), \quad (1)$$

где $\hat{\tau}$ — оценка времени задержки, а $R_{a_1,a_2}(\tau)$ — ФКК временных огибающих основного a_1 и опорного a_2 сигналов.

Ключевым в оценке ВС (1) является вычисление огибающих. В настоящей работе для этого используется модифицированная процедура „выпрямление и фильтрация“ [12]. Обозначим дискретный временной сигнал как $x(i)$, где i — временной индекс, тогда его огибающая $a(i)$ может быть получена как:

$$a(i) = \text{ФВЧ}(\text{ФНЧ}(|x(i)|)), \quad (2)$$

где $|\cdot|$ — символ вычисления абсолютной величины (т.е. „выпрямления“) сигнала, а ФНЧ и ФВЧ — фильтры низких и высоких частот соответственно.

Фильтр низких частот предназначен для сглаживания выпрямленного сигнала и устранения выбросов. Сглаживание осуществляется фильтром первого порядка [13]:

$$y(i) = \beta(x(i) + x(i-1)) + \alpha y(i-1) \ , \quad (3)$$

где $x(i)$ и $y(i)$ — входной и выходной сигналы фильтра. Коэффициент α ($0 \le \alpha < 1$) задается на основе соотношения:

$$\alpha = 1 - 2/(1 + T_{\text{нч}} F_s), \quad (4)$$

где F_s — частота дискретизации сигнала в герцах, а $T_{\text{нч}}$ — эквивалентная длина окна в секундах, $\beta = (1-\alpha)/2$. Величина $T_{\text{нч}}$ должна соотноситься с темпом модуляции акустических сигналов (речи, музыки). Если значение $T_{\text{нч}}$ мало, то полученная огибающая будет сильно флуктуировать, если велико, это приведет к сильному сглаживанию самих огибающих. В обоих случаях уменьшится корреляция между огибающими и соответственно снизится точность оценки ВС. Таким образом, существует некая оптимальная длина окна $T^*_{\text{нч}}$.

Фильтр высоких частот предназначен для удаления постоянной и низкочастотных составляющих сглаженных огибающих. ВЧ-фильтрация также осуществляется фильтром первого порядка [13]:

$$y(i) = \eta(x(i) - x(i-1)) + \gamma y(i-1) \ , \quad (5)$$

где $\gamma = 1 - 2/(1 - T_{\text{вч}} F_s)$, а $\eta = (1+\gamma)/2$. ВЧ-фильтрация приводит, с одной стороны, к уменьшению корреляции огибающих, а с другой — к сужению главного лепестка ФКО, т.е. можно предположить, что также существует некое оптимальное $T^*_{\text{вч}}$ (заметим, что $T^*_{\text{вч}}$ и $T^*_{\text{нч}}$ в общем случае различны).

В качестве примера на рис. 2 представлены отрезок речевого сигнала (*1*), его огибающая после сглаживания (*2*) и после ВЧ фильтрации (*3*).

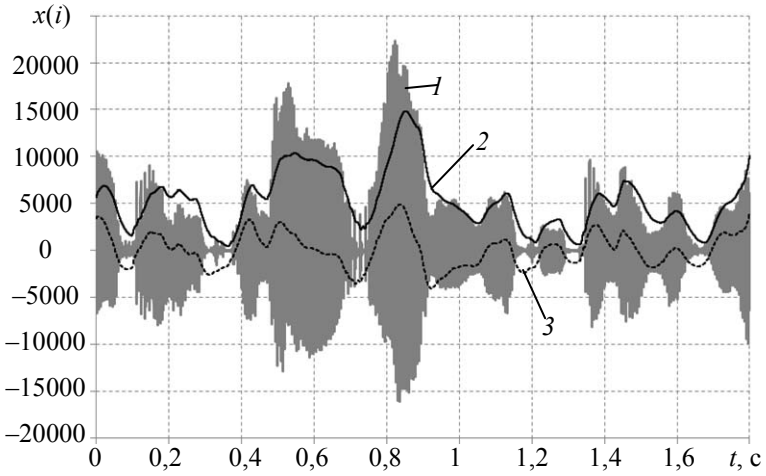


Рис. 2

Функция кросскорреляции. Качество оценки ФКО зависит от длины блока анализа данных T_a. Величина T_a должна соответствовать периодам осцилляции огибающих аудиосигнала. Если T_a < 0,1 с, то огибающая речевого сигнала может представлять собой монотонно возрастающую или убывающую функцию. В этом случае невозможно корректно оценить ФКО. Поскольку основная часть спектра огибающих расположена на частоте 4 Гц и выше, то адекватные оценки ВС получаются при $T_a \geq$ 1—2 с.

Поскольку вычисление ФКО на таких интервалах требует существенных вычислительных затрат, то вместо стандартной формулы вычисления ФКК [14]:

$$R_{x_1,x_2}(m) = \frac{\sum_i \left((x_1(i)-\bar{x}_1)(x_2(i)-\bar{x}_2)\right)}{\sqrt{\sum_i \left((x_1(i)-\bar{x}_1)^2\right)\sum_i \left((x_2(i)-\bar{x}_2)^2\right)}}$$

используем вычисление за один цикл и „с шагами“, значительно ускорив процесс без потери точности:

$$R_{x1,x2}(m) = \frac{\sum_i x_1(Ki)x_2(Ki-m) - \frac{1}{M}\left(\sum_i x_1(Ki)\sum_i x_2(Ki-m)\right)}{\sqrt{d}}, \qquad (6)$$

$$d = \left(\sum_i x_1^2(Ki) - \frac{\left(\sum_i x_1(Ki)\right)^2}{M}\right)\left(\sum_i x_2^2(Ki-m) - \frac{\left(\sum_i x_2(Ki-m)\right)^2}{M}\right). \qquad (7)$$

Здесь $K > 0$ — шаг вычисления; $x_1(i)$ и $x_2(i)$ — дискретные сигналы; N — полное число отсчетов в сигналах на блоке анализа; $m = 0, \pm 1, \pm 2, ...$ — временная задержка; $\bar{x}$ — среднее значение; $M = \lfloor (N-m)/K \rfloor$ — количество отсчетов огибающих в вычислении каждого из значений ФКО; $i = 0, ..., M-1$; $\lfloor\ \rfloor$ — символ „взятие целой части“.

Поскольку огибающая речевого сигнала осциллирует медленно, то можно задавать шаг вычисления K значительно больше единицы, что существенно ускоряет вычисления. Так как основная часть модуляционных компонент огибающих аудиосигналов находится в диапазоне до 25 Гц [15], то должно быть $K < 0{,}5F_s/25$. Для сигналов F_s=16 кГц было принято $K = 100$.

Пример ФКО реальных записей музыкальных сигналов и их огибающих представлен на рис. 3. Цифровой опорный сигнал воспроизводился через аудиоколонку. Основной сигнал был записан через микрофон в помещении с временем реверберации 650 мс, расстояние между громкоговорителем и микрофоном равнялось 4 м. Искажения основного сигнала трактом воспроизведения и реверберацией привели к тому, что корреляция между сигналами мала (кривая *1* — значение максимума, помеченное кружком, при $\tau = 0$ равно 0,11). С другой стороны, видно, что корреляция как огибающих (*2*), так и огибающих после ВЧ-фильтрации (*3*) существенна.

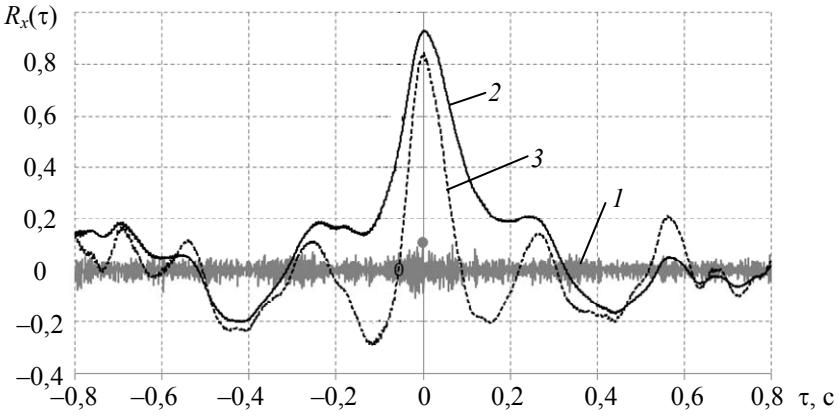


Рис. 3

Исследование влияния искажений сигналов на оценку ФКО. Пусть $x_1(i)$ и $x_2(i)$ — дискретные временные сигналы с нулевыми средними. Обозначим $R_{x_1,x_2}(m)$ — ФКК, $R_{a_1,a_2}(m)$ — ФКО сигналов. Если $x_1(i) = x_2(i)$, то $\hat{R}_{x_1,x_2}(0) = \hat{R}_{a_1,a_2}(0) = 1$ (здесь $\hat{R}$ — оценка R).

Нелинейные преобразования. Рассмотрим простые нелинейные преобразования: $x_2(i) => |x_2(i)|$, или $x_2(i) => (x_2(i))^2$. Можно показать, что в этом случае значение $R_{x_1,x_2}(0)$ существенно снижается, в то время как $R_{a_1,a_2}(0)$ меняется незначительно.

Исследование влияния шума. Зададим $x_1(i)$ и $x_2(i)$:

$$x_1(i) = (1-\mu)s(i) + \mu n_1(i), \quad x_2(i) = (1-\mu)s(i) + \mu n_2(i), \tag{8}$$

где $s(i)$ — речевой сигнал; $n_1(i)$ и $n_2(i)$ — последовательности независимых случайных величин, $0 \le \mu \le 1$. При $\mu = 0$ $x_1(i) = x_2(i) = s(i)$ и $\hat{R}_{x_1,x_2}(0) = \hat{R}_{a_1,a_2}(0) = 1$. При $\mu = 1$ $x_1(i)$ и $x_2(i)$ являются исходными независимыми случайными величинами и $\hat{R}_{x_1,x_2}(0) \approx 0$ и $\hat{R}_{a_1,a_2}(0) \approx 0$. Если дисперсии $s(i)$, $n_1(i)$ и $n_2(i)$ равны, то получим теоретические выражения для $R_{x_1,x_2}(0)$ как функцию от μ:

$$R^t_{x_1,x_2}(0,\mu) = \frac{(1-\mu)^2}{\mu^2 + (1-\mu)^2}. \tag{9}$$

На рис. 4 приведены оценки $\hat{R}_{x_1,x_2}(0)$, $\hat{R}_{a_1,a_2}(0)$, их 95 %-ные доверительные интервалы для сигналов (8) как функция от μ. Речевые сигналы брались из базы TIMIT [16], в качестве шума был взят файл factory1.wav из базы NOISEX-92 [17]. Мощности сигналов речи и шума приводились к единой величине перед преобразованием (8). Параметры вычисления огибающих: $T_a = 2$ с, $T_{\text{нч}} = 0,05$ с, ВЧ-фильтр не использовался. Полученные результаты показывают,

что при увеличении доли шума $\hat{R}_{x_1,x_2}(0)$ (кривая *1*) уменьшается, почти совпадая с теоретической кривой *3*, в то же время $\hat{R}_{a_1,a_2}(0)$ (кривая *2*) сохраняет достаточно высокие значения вплоть до $\mu = 0,6$.

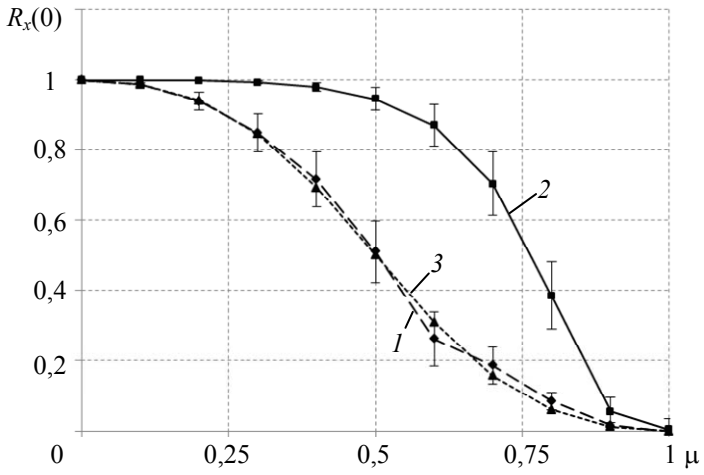


Рис. 4

Эксперименты: поиск оптимальных параметров алгоритма. Целью экспериментов являлся выбор оптимальных значений постоянных времени $T^*_{\text{нч}}$ и $T^*_{\text{вч}}$ ФНЧ и ФВЧ для различных T_a. Использовались двухканальные записи сигналов: „речь“, „песня“, „музыка“, „розовый шум“ и „модулированный по амплитуде белый шум“, записанные в помещении с постоянной времени реверберации 650 мс. Расстояние между основным и опорным микрофонами 4 м, соответственно теоретически рассчитанная задержка между сигналами для частоты дискретизации 16 кГц равнялась 183 отсчетам. В качестве целевой величины был выбран средний квадрат ошибки (mean squared error, MSE) оценки ВС:

$$\text{MSE}(\tau) = \frac{1}{L}\sum_{i=0}^{L-1}\left(\tau(i) - \tau_{\text{teor}}\right)^2,$$

где L — общее число экспериментов по оценке задержки; τ_{teor} — теоретическое значение задержки. Оптимальные значения параметров, полученные экспериментально, приведены в таблице.

T_a, с	ФНЧ	ФНЧ+ФВЧ	
	$T^*_{\text{нч}}$, с	$T^*_{\text{нч}}$, с	$T^*_{\text{вч}}$, с
2	0,0212	0,0396	0,0319
3	0,0219	0,0311	0,0441
4	0,0241	0,0313	0,0394
5	0,0107	0,0275	0,0332
6	0,0119	0,0315	0,0327
7	0,0102	0,0303	0,0275
8	0,0137	0,0225	0,0374
Среднее	0,0164	0,0321	0,0340

Сравнение МКО с другими методами оценки ВС. Предложенный метод сравнивался с кросскорреляционным и методом PHAT.

Через аудиоколонку проигрывалась музыка, записанная на компакт-диске, сигнал с которого использовался в качестве опорного, основной записывался через удаленный микрофон

в помещении и представлял собой сумму речевого сигнала и проигрываемой музыки.

Экспериментальные исследования показали, что в случаях, когда искажения основного и опорного сигналов невелики, лучшие результаты дает РНАТ (*1*), средние — ФКК (*2*), а предложенный метод (*3*) неэффективен. Однако если сигналы сильно искажены, МКО дает лучшие результаты — минимальное MSE (рис. 5).

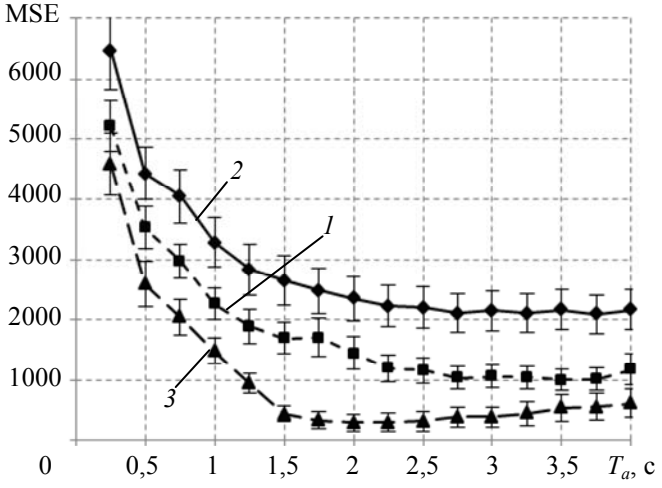


Рис. 5

Обсуждение. Полученные в работе результаты позволяют утверждать, что использование временных огибающих речевых сигналов в задаче оценки временного сдвига между аудиосигналами оправдано в случаях, когда искажения сигналов слабо влияют на огибающие. Например, МКО полезен при асинхронной фильтрации речевых сигналов [9].

Традиционные методы оценки ВС эффективнее метода МКО в случае слабых искажений самих сигналов или в случае, когда огибающие имеют сильную не меняющуюся периодичность (например, на сигналах типа „ритмичная музыка“).

По нашему мнению, вопрос выбора параметров $T_{\text{нч}}$ и $T_{\text{вч}}$ остается открытым. Эти параметры, как показывает моделирование, в общем случае зависят от характеристик как сигнала, так и его искажений. Однако соответствие полученных результатов обобщенным характеристикам спектра огибающих речевых сигналов позволяет предположить, что данные таблицы могут служить первым приближением для реальных параметров обработки.

Заключение. В работе описан и исследован метод оценки временного сдвига между двумя акустическими сигналами, основанный на кросскорреляции их огибающих. Главным достоинством метода является то, что он показывает хорошие результаты в случаях сильных искажений сигналов, например, при реверберации, или в асинхронном случае, когда сигналы записывались в разных условиях на разной аппаратуре. Недостатком является большая длина блоков данных, необходимых для оценки ВС.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Chen J., Benesty J., Huang Y. A.* Time Delay Estimation in Room Acoustic Environments // EURASIP J. on Advances in Signal Processing. 2006. P. 1—20.

2. *Sandmair A., Lietz M., Stefan J., Leon F. P.* Time delay estimation in the time-frequency domain based on a line detection approach // Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Prague, Czech Republic, 2011. P. 2716—2719.

3. *Gedalyahu K., Eldar Y. C.* Time-delay estimation from low-rate samples: A union of subspaces approach // IEEE Transactions on Signal Processing. 2010. Vol. 58. N 6. P. 3017—3031.

4. *Kirkwood B.* Acoustic Source Localization Using Time-Delay Estimation: M.S. Thesis. Technical University of Denmark, 2003.

5. *Kozlov A., Kudashev O., Matveev Yu., Pekhovsky T., Simonchik K., Shulipa A.* SVID Speaker Recognition System for NIST SRE 2012 // Proc. of 15th Intern. Conf. "Speech and Computer" (SPECOM 2013). Springer Lecture Notes in Computer Science. Lecture Notes in Artificial Intelligence. 2013. Vol. 8113. P. 278—285.

6. *Bédard S., Champagne B., Stéphenne A.* Effects of Room Reverberation on Time-Delay Estimation Performance // Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Adelaide, SA, 1994. Vol. 2. P. 261—264.

7. *Raya R., Frizera A., Ceres R., Calderón L., Rocon E.* Design and evaluation of a fast model-based algorithm for ultrasonic range measurements // Sensors and Actuators A: Physical. 2008. Vol. 148, N 1. P. 335—341.

8. *Yang L., Lavrinenko A.V., Hvam J.M., Sigmund O.* Design of one-dimensional optical pulse-shaping filters by time-domain topology optimization // Appl. Phys. Lett. 2009. Vol. 95, Is. 26. P. 261 101.

9. *Алейник С. В., Столбов М. Б.* Подавление акустических помех аудиоустройств с использованием асинхронного опорного сигнала // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 11—18.

10. *Lazarov B. S., Matzen R., Elesin Y.* Topology optimization of pulse shaping filters using the Hilbert transform envelope extraction // Structural and Multidisciplinary Optimization. 2011. Vol. 44, N 3. P. 409—419.

11. *Thrane N., Wismer J., Konstantin-Hansen H., Gade S.* // Application Note. Practical use of the Hilbert transform. Techn. rev. N 3. [Электронный ресурс]: <http://www.bksv.com/doc/bo0437.pdf>.

12. *Bouzid O. M., Tian G. Y., Neasham J., Sharif B.* Envelope and Wavelet Transform for Sound Localisation at Low Sampling Rates in Wireless Sensor Networks // J. of Sensors. 2012. Vol. 2012. P. 680 383.

13. *Orfanidis S. J.* Introduction to Signal Processing. [Электронный ресурс]: <http://www.ece.rutgers.edu/~orfanidi/intro2sp/orfanidis-i2sp.pdf>.

14. *Aarts R. M., Irwan R., Janssen A. J. E. M.* Efficient tracking of the cross-correlation coefficient // IEEE Transact. on Speech and Audio Processing. 2002. Vol. 10, N 6. P. 391—402.

15. *Hougast T., Steeneken H. J. M.* A review of the MTF concept in room acoustics and it's use for estimating speech intelligibility in auditoria // J. of the Acoustical Society of America. 1985. Vol. 77, Is. 3. P. 1069—1077.

16. TIMIT Acoustic-Phonetic Continuous Speech Corpus. [Электронный ресурс]: <http://catalog.ldc.upenn.edu/LDC93S1>.

17. Database of recording of various noises NOISEX-92 [Электронный ресурс]: <http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html>.

Сведения об авторах

Сергей Владимирович Алейник — ООО „ЦРТ-инновации", Санкт-Петербург; научный сотрудник; E-mail: aleinik@speechpro.com

Михаил Борисович Столбов — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ-инновации", Санкт-Петербург; старший научный сотрудник; E-mail: stolbov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 621.391.037.372

С. В. Алейник, М. Б. Столбов

СТОХАСТИЧНОСТЬ РЕЧЕВЫХ СИГНАЛОВ И ЕЕ ОЦЕНИВАНИЕ

Проанализированы способы оценки стохастичности речевых сигналов. Результаты моделирования показали, что предложенный способ оценки превосходит известные по качеству — оценки имеют меньшее смещение и дисперсию.

Ключевые слова: *стохастичность, речевые сигналы, адаптивная фильтрация, предсказание.*

Введение. В различных областях научно-технической деятельности при анализе временных рядов часто используются такие понятия, как „случайность“ и „стохастичность“ [1—9]. Существуют различные подходы, позволяющие численно охарактеризовать данные понятия, т.е. получить числовые значения неких параметров временных рядов. Применительно к анализу речевых сигналов наиболее используемыми параметрами, характеризующими случайность, являются спектральная энтропия (spectral entropy, SE) [3, 4], так называемая спектральная „ровность“ или „плоскостность“ (spectral flatness, SF) [5, 8] и частота переходов через нуль (zero-crossing rate, ZCR) [6, 7]. Эти параметры обычно используются для разделения звуков на звонкие/глухие, определения пауз, детектирования речи и т.д. Например, в работах [3, 4] указывается, что SE может использоваться как некий признак в задаче распознавания речи, в [5] описано применение SF для разделения дикторов, в [6, 7] ZCR используется для детектирования глухих звуков. Сравнительный анализ SF и SE приведен в статье [8].

Несмотря на то что SF и SE вычисляются по-разному, оба параметра характеризуют одно физическое свойство сигнала — его стохастичность [1—9]. Если анализируемый сигнал — гармоническое колебание или постоянная составляющая, то его оцененные SF и SE равны нулю [8]. С другой стороны, если сигнал — белый шум, то оценки его SF и SE флуктуируют около некоей ненулевой положительной величины (около единицы в случае нормализации). Иногда ZCR также считается мерой стохастичности речевого сигнала [6]. В работе [10] автором был предложен подход к оценке стохастичности, основанный на использовании адаптивного линейного предсказателя (adaptive linear predictor, ALP [11]).

Целью настоящей работы является анализ способов измерения стохастичности, а также исследование нового способа оценки стохастичности речевых сигналов.

Стохастичность и ее измерение. Характеристики процессов, применяемые в теории хаоса (например, экспоненты Херста и Ляпунова [12, 13] и т.п.), зачастую требуют больших объемов данных и в основном описывают систему или процесс в целом. При анализе речевых сигналов требуется оценивать динамику процесса на коротких интервалах времени. В настоящей статье примем, согласно [2, 14], что стохастичность — значит случайность, а случайность подразумевает отсутствие предсказуемости. Подобный подход (некорректный с точки зрения специалистов, занимающихся, например, анализом хаотических систем) вполне применим в анализе речевых сигналов; используя его, будем считать, что:

1) белый шум (полностью непредсказуемый процесс) есть „полностью“ стохастический процесс со стохастичностью теоретически равной единице;

2) постоянная составляющая, или гармонический процесс (полностью предсказуемые сигналы) являются детерминированными сигналами со стохастичностью, теоретически равной нулю;

3) розовый шум, сумма белого шума и, например, гармонического сигнала, а также хаотические процессы теоретически имеют стохастичность между нулем и единицей.

Данный подход показывает, что SE и SF не совсем корректно применять для оценки стохастичности [10]. В самом деле, физически SE и SF являются мерами „узкополосности“ спектра. Если спектр мощности на всех частотах, кроме одной, равен нулю, то как SE, так и SF, рассчитанные по такому спектру, равны нулю. При увеличении широкополосности спектра соответствующие SE и SF увеличиваются. Конечно, гармонический сигнал имеет узкополосный, а белый шум — широкополосный спектр, это приводит к малым и большим значениям соответствующих SE и SF. Однако оценки спектра мощности белого шума сильно флуктуируют, что приводит к значительному смещению вниз рассчитанных SE и SF. С другой стороны, последовательность прямоугольных импульсов (детерминированный сигнал) в спектральной области представляет собой набор гармоник, поэтому вычисленное по подобному спектру значение SE оказалось ненулевым. ZCR может рассматриваться лишь как грубая оценка средней частоты сигнала [6], а не его стохастичности. Пусть, например, x_i $(i = 0, N-1)$ есть дискретный сигнал с нулевым средним. Тогда легко показать, что среднее значение ZCR=0,5 (вычисленное, например, с использованием (1) в [6]) в случае, если x_i — белый шум. С другой стороны:

$$\mathrm{ZCR} = \begin{cases} 0, & \text{если } x_i = c, \\ 1, & \text{если } x_i = +c, -c, +c, \ldots = (-1)^i c, \quad \forall i = 0,1,\ldots, \end{cases} \tag{1}$$

где c — постоянная величина. Ясно, что $x_i = (-1)^i c$ является не стохастическим, а детерминированным сигналом с максимально возможной дискретной частотой. Если, например, пропустить белый шум через высокочастотный фильтр, то ZCR результирующего сигнала окажется выше, чем ZCR исходного шума, в то время как из физических соображений понятно, что стохастичность должна уменьшиться.

С другой стороны, алгоритмы линейного предсказания, широко применяющиеся, например, в адаптивной фильтрации и кодировании речи, позволяют непосредственно вычислять предсказуемость сигнала и более корректно измерять его стохастичность.

Основная идея подхода. В работе [11] был описан и исследован коэффициент стохастичности, вычисляемый как:

$$S_P = <\xi^2> \big/ <x^2> = \sum_i \xi_i^2 \Big/ \sum_i x_i^2 , \tag{2}$$

где $<>$ — символ осреднения по времени, x_i — входной сигнал и ξ_i — выходной сигнал ошибки адаптивного линейного предсказателя (ALP) [11].

В работе [11] показано, что:

1) если x_i — детерминированный процесс, то ξ_i=0, $S_P = 0$;

2) если x_i — белый шум, то $<\xi^2> \approx <x^2>$, $E(S_P) = 1$ (здесь $E(\cdot)$ — символ математического ожидания: в реальности $S_P \approx 1$ из-за флуктуаций $<\xi^2>$ и $<x^2>$).

Параметр S_P характеризует непредсказуемость сигнала [10], и если x_i можно представить как сумму его детерминированной и стохастичной компонент, то:

$$E(S_P) = P_s / P = P_s / (P_s + P_d), \tag{3}$$

где P — мощность сигнала, а P_d и P_s — мощность его детерминированной и стохастичной составляющих соответственно.

Детальный анализ (2) показывает, что S_P есть величина, обратная к “prediction gain” (PG) — так называемый выигрыш, усиление предсказания, используемая в кодировании речи для оценки качества предсказателей [15]. Таким образом, возможно измерить S_P не с помощью

ALP, а с помощью других алгоритмов линейного предсказания, например, можно с помощью рекурсии Левинсона—Дурбина [16] оценить величину PG и затем вычислить:

$$S_G = 1/\mathrm{PG}\,. \tag{4}$$

Предлагаемый коэффициент стохастичности. Отметим, что ALP имеет в общем случае выход сигнала ошибки ξ_i и выход предсказанного сигнала y_i [11]. После завершения процесса адаптации в идеальном случае на выходе сигнала ошибки присутствует только стохастическая компонента входного сигнала, а на выходе предсказанного — только детерминированная [11], т.е. по аналогии с (2), используя не сигнал ошибки, а предсказанный сигнал y_i, получим

$$S = < y^2 > / < x^2 > = \sum_i y_i^2 \Big/ \sum_i x_i^2\,, \tag{5}$$

откуда

$$E(S) = P_d / (P_s + P_d)\,, \tag{6}$$

т.е. $E(S) = 1$, когда x_i — детерминированный процесс, $E(S) = 0$, когда x_i — полностью стохастический. Сравнив (3) и (6), получим:

$$E(S_P) = 1 - E(S)\,, \tag{7}$$

следовательно, ALP позволяет оценивать стохастичность сигнала как S_P и как $1 - S$. Одним из базовых свойств ALP является то, что корреляция между сигналом ошибки ξ_i и предсказанным сигналом y_i мала (равна нулю в случае идеального разделения входного сигнала на стохастическую и детерминированную составляющие) [11]. Соответственно можно утверждать, что среднее арифметическое:

$$\overline{S} = 0{,}5(S_P + (1 - S)) \tag{8}$$

будет иметь меньшую дисперсию, чем S_P и S.

Используя (2), (5), (8), а также выражение для сигнала ошибки: $\xi_i = x_i - y_i$ из [11], получим формулу для нового коэффициента стохастичности:

$$\overline{S} = S_C = \frac{1}{2}\left(1 + \frac{\sum_i \xi_i^2}{\sum_i x_i^2} - \frac{\sum_i y_i^2}{\sum_i x_i^2}\right) = \frac{1}{2}\left(1 + \frac{\sum_i (x_i - y_i)^2 - \sum_i y_i^2}{\sum_i x_i^2}\right) = \frac{1}{2}\left(1 + \frac{\sum_i x_i^2 - 2\sum_i x_i y_i}{\sum_i x_i^2}\right) = 1 - \frac{\sum_i x_i\, y_i}{\sum_i x_i^2}\,. \tag{9}$$

Легко показать, что значение S_C, так же как и S_P, изменяется от 0 до 1 и характеризует „непредсказуемость“ входного сигнала x_i: чем более „непредсказуем“ сигнал, тем ближе S_C к 1. Действительно, величина $\sum_i x_i\, y_i$ в (9) есть не что иное, как оценка ненормированного корреляционного коэффициента входного и предсказанного сигналов, она близка к нулю, когда x_i — стохастический сигнал (например, белый шум), в этом случае S_C близок к единице. С другой стороны, если x_i — детерминированный сигнал (например, синусоидальный), то $y_i = x_i$ после завершения адаптации, и тогда $\dfrac{\sum_i x_i\, y_i}{\sum_i x_i^2} = 1$, а $S_C = 0$.

Следует заметить, что оценки как S_C, так и S_P могут принимать значения больше единицы из-за флуктуаций и некорректно выбранных параметров. Данный феномен будет рассмотрен ниже.

Моделирование. Характеристики коэффициентов стохастичности SE, S_G, S_P и S_C исследовались с помощью статистического моделирования. Во всех экспериментах использовались следующие параметры: длина блока анализа K=512 отсчетов; число M коэффициентов

линейного и адаптивного линейного предсказателей LPC и ALP равно 16; количество испытаний для оценки распределений и средних значений равно 10^6. С целью приведения области определения всех коэффициентов к единому интервалу [0, 1] в качестве SE использовалась нормированная энтропия спектра [10]:

$$\mathrm{SE} = -1/\log_2(N) \sum_{i=0}^{N-1} X_i \log_2(X_i), \qquad (10)$$

где N — количество отсчетов спектра мощности; X_i — величина i-го спектрального отсчета, нормализованная так, что: $\sum_{i=0}^{N-1} X_i = 1$. Спектральные окна, перекрытия и накопления не использовались. Известный алгоритм Левинсона—Дурбина [16] использовался для получения коэффициентов LPC и величины PG, которая затем пересчитывалась в S_G (4). В качестве ALP использовался нерекурсивный адаптивный сумматор [11], который адаптировался по нормализованному методу наименьших квадратов (normalized least-mean squares algorithm, NLMS) с постоянной адаптации $\mu = 0{,}1$ и регуляризационной константой $\delta = 1$ [17].

Плотности распределения и средние значения. На рис. 1 представлены экспериментально определенные плотности распределения коэффициентов SE (*1*), S_G (*2*), S_P (*3*) и S_C (*4*) для входных процессов различной стохастичности (*а* — плотность распределения для белого шума, теоретическое значение стохастичности равно 1).

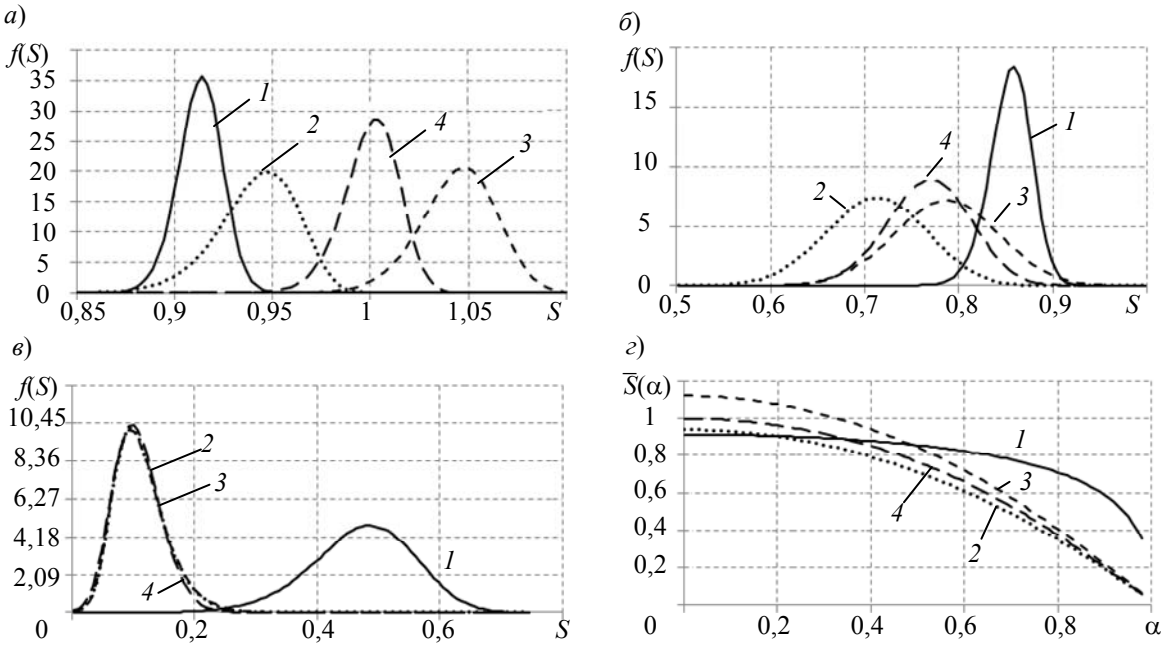


Рис. 1

На рис. 1, *а* видно заметное смещение SE и S_G вниз, причем смещение SE максимально. Этот факт легко объяснить, поскольку SE=1 только тогда, когда $X_i = \mathrm{const}$ для всех $i = 0, N-1$. В действительности каждый отсчет оцененного спектра мощности достаточно сильно флуктуирует. С другой стороны, S_P смещен вверх — видно, что его среднее больше 1. По нашему мнению, это вызвано тем, что процесс адаптации не успел завершиться на длине блока сигнала. Данный эффект зависит от самой длины и постоянной μ. Из рис. 1, *а* также видно, что предлагаемый коэффициент S_C имеет наименьшее смещение и флуктуирует около теоретического значения 1. На рис. 1, *б* и *в* приведены распределения коэффициентов стохастичности розового шума, сгенерированного с помощью фильтра низких частот первого порядка [18]:

$$y_i = \alpha y_{i-1} + (1-\alpha)x_i, \qquad (11)$$

где y_i и x_i — розовый и белый шум в момент времени i, а $0 \le \alpha < 1$ — параметр фильтра. Известно, что чем ближе α к 1, тем более предсказуемым (и менее стохастичным) становится y_i, т.е. на рис. 1 изображены результаты: для белого шума (*а* — $\alpha = 0$), для среднестохастичного процесса (*б* — 0,5) и для низкостохастичного процесса (*в* — 0,95). Видно, что средние значения всех коэффициентов уменьшаются при увеличении α; в случае низкостохастичного процесса (рис. 1, *в*) все коэффициенты, исключая SE, имеют практически идентичные распределения; SE низкостохастичного процесса имеет наибольшую дисперсию и смещена вверх относительно других коэффициентов. Легко понять также, что чувствительность к изменению α различна для различных коэффициентов. Данный факт подтверждает рис. 1, *г*, где приведены средние значения всех коэффициентов как функции от $\alpha = [0; 0{,}99]$.

Устойчивость к изменению параметров. Несмотря на большое смещение и дисперсию коэффициент SE удобно использовать. Во-первых, для вычисления SE требуется задать только длину блока анализа. И во-вторых, величина SE всегда стабильна и дает адекватную (хоть и смещенную) оценку стохастичности. Для вычисления S_G следует задать длину блока анализа и количество LPC коэффициентов; для S_P и S_C — длину блока анализа, количество ALP коэффициентов, постоянную адаптации $\mu > 0$ и регуляризационную константу δ.

Чем больше длина блока анализа, тем меньше дисперсия соответствующего коэффициента, значение δ также не оказывает большого влияния на S_P и S_C (если только оно не экстремально мало или велико); значение $\delta = 1$ допустимо для большинства реальных ситуаций. В противоположность этому число коэффициентов LPC и ALP, так же как и μ, — важные параметры. Результаты эксперимента по определению среднего значения S_G (*1*), S_P (*2*) и S_C (*3*) белого шума при различном числе коэффициентов LPC и ALP (*M*) представлены на рис. 2. Видно, что увеличение *M* вызывает заметное снижение S_G. По нашему мнению, данный эффект необходимо принять во внимание, кроме того моделирование показывает, что эффект наиболее сильно выражен для розового шума с высоким уровнем α и других низкостохастичных сигналов.

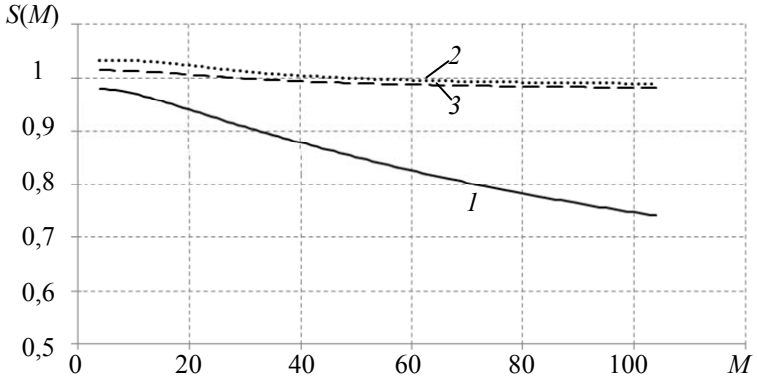


Рис. 2

Наиболее критичным параметром для вычисления S_P и S_C является μ. Известно [19], что теоретическими границами для устойчивой работы алгоритма NLMS являются $0 < \mu < 2$. При $\mu \ge 2$ адаптация в ALP становится нестабильной, что вызывает резкое неадекватное увеличение как среднего значения S_P и S_C так и их дисперсии. С другой стороны, в случае некорректно малого μ (рис. 3, *а*, область $\mu < 0{,}05$) процесс адаптации идет медленно и не

может завершиться в момент окончания обработки блока данных, что также вызывает увеличение дисперсии (*1* — K=64, *2* — 128, *3* — 256, *4* — 512). Среднеквадратические отклонения (RMSE) для S_P и S_C как функций от μ показаны на рис. 3, *б*. Для улучшения восприятия графики разбиты на две части: слева — $0 < \mu < 1$ и справа — $1 < \mu < 2$ (левая и правые вертикальные оси соответственно). Видно, во-первых, резкое увеличение RMSE при $\mu > 1$, во-вторых, RMSE S_C (*1*) всегда меньше, чем у S_P (*2*) — т.е. предлагаемый коэффициент стабильнее даже при неоправданно высоком μ.

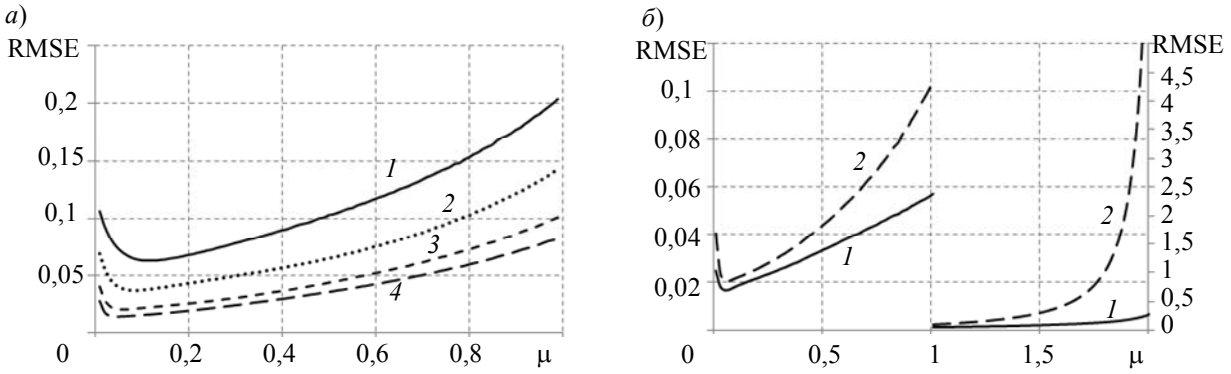


Рис. 3

Практические рекомендации. Значения рассмотренных коэффициентов не зависят от мощности сигнала, но постоянная составляющая сигнала или помеха даже малой мощности могут затруднять интерпретацию результатов. Например, при обработке речи шум слабого уровня в паузах между звуками вызывает увеличение значения коэффициентов, постоянная составляющая, наоборот — уменьшение. Поэтому желательно при предобработке удалять постоянную составляющую сигнала; добавлять некое фиксированное смещение к сигналу (обычно для реальных речевых сигналов просто прибавлять к сигналу величину 50—500).

Также следует отметить, что алгоритмы адаптивной фильтрации в ALP хорошо подходят для поотсчетной обработки сигналов — в этом случае в (9) суммирование по блоку можно заменить экспоненциальным усреднением или какой-либо другой НЧ-фильтрацией.

Применение и пример обработки речевого сигнала. На практике коэффициент S_C может применяться как характеристика сигнала в различных задачах, например, для классификации звуков. Значение S_C высоко (возрастает) на фрикативных согласных: ‘s’, ‘sh’, ‘ch’ и т.п., а также на коротких взрывных согласных, низко на гласных и звонких согласных. Например, данное свойство проиллюстрировано на рис. 4, где для известной фразы из базы TIMIT [20] “She had your dark suit in greasy wash water all year” представлен вычисленный S_C.

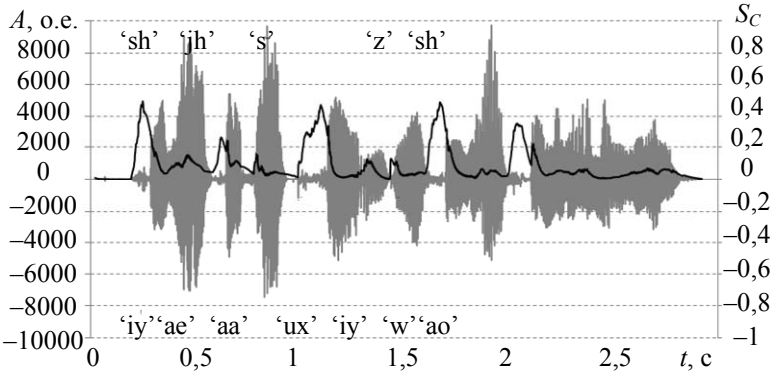


Рис. 4

Заключение. В работе представлен простой подход к вычислению стохастичности речевых сигналов, имеющий преимущества (меньшие смещение и дисперсия и более высокая устойчивость к изменению параметров обработки) по сравнению с известными способами. Результаты моделирования подтверждают эффективность подхода при обработке, например, речевых сигналов для классификации звуков.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Puente C. E., Obregón N., Sivakumar B.* Chaos and stochasticity in deterministically generated multifractal measures // Fractals. 2002. Vol. 10, N 1. P. 91—102.
2. *Sivakumar B.* Is a Chaotic Multi-Fractal Approach for Rainfall Possible? // Hydrological Processes. 2001. Vol. 15, N 6. P. 943—955.
3. *Misra H., Ikbal S., Sivadas S., Bourlard H.* Multi-resolution Spectral Entropy Feature for Robust ASR // Proc. of IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Philadelphia, USA, 2005. Vol. 1. P. 253—256.
4. *Toh A. M., Togneri R., Nordholm S.* Spectral entropy as speech features for speech recognition // Proc. of 6th Postgraduate Electrical Engineering & Computing Symposium (PEECS). 2005. P. 22—25.
5. *Bardeli R.* Source Separation Using the Spectral Flatness Measure // Proc. of CHiME 2011 Workshop on Machine Listening in Multisource Environments. 2011. P. 80—85.
6. *Bachu R. G., Kopparthi S., Adapa B., Barkana B. D.* Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal // Proc. American Society for Engineering Education. 2008. P. 1—7.
7. *Khan A. U., Bhaiya L. P., Banchhor S. K.* Hindi Speaking Person Identification Using Zero Crossing Rate // Intern. J. of Soft Computing and Engineering. 2012. Vol. 2, N 3. P. 101—104.
8. *Madhu N.* Note on Measures for Spectral Flatness // Electronics Letters. 2009. Vol. 45, Is. 23. P. 1195—1196.
9. *Dubnov S.* Non-gaussian source-filter and independent components generalizations of spectral flatness measure // Proc. of 4th Intern. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003). Nara, Japan, 2003. P. 143—148.
10. *Алейник С.* Оценка детерминированности временных рядов // Радиотехника. 1999. № 9. С. 16—22.
11. *Widrow B., Lehr M., Beaufays F., Wan E., Bileillo M.* Learning algorithms for adaptive processing and control // IEEE Intern. Conf. on Neural Networks. San Francisco, CA, USA, 1993. Vol. 1. P. 1—8.
12. *Delignieres D., Ramdani S., Lemoine L., Torre K., Fortes M., Ninot G.* Fractal analyses for short time series: A re-assessment of classical methods // J. of Mathematical Psychology. 2006. Vol. 50, N 6. P. 525—544.
13. *Wolf A., Swift J., Swinney H., Vastano J.* Determining lyapunov exponents from a time series // Physica D: Nonlinear Phenomena. 1985. Vol. 16, N 3. P. 285—317.
14. Randomness [Электронный ресурс]: <http://encyclopedia.thefreedictionary.com/stochasticity>.
15. *Heim A., Sorger U., Hug F.* Doppler-variant modeling of the vocal tract // Proc. of IEEE Intern/ Conf/ on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, Nevada, USA, 2008. P. 4197—4200.
16. *Corneliu M., Costinescu B.* Implementing the Levinson-Durbin Algorithm on the SC140 // Freescale Semiconductor. Application Note. 2005. AN2197. Rev. 1. 24 p.
17. *Bitzer J., Brandt M.* Speech Enhancement by Adaptive Noise Cancellation: Problems, Algorithms and Limits // Proc. of 39th Intern. Conf. Audio Forensics: Practices and Challenges (AES-39). Denmark, 2010. P. 106—113.
18. *Orfanidis S. J.* Introduction to Signal Processing. [Электронный ресурс]: <http://www.ece.rutgers.edu/~orfanidi/intro2sp/orfanidis-i2sp.pdf>.
19. *Haykin S.* Adaptive Filter Theory. Prentice-Hall, 1996.
20. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT). Training and Test Data. NIST Speech Disc CD1-1.1 [Электронный ресурс]: <http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html>.

Сведения об авторах

Сергей Владимирович Алейник — ООО „ЦРТ-инновации“, Санкт-Петербург; научный сотрудник; E-mail: aleinik@speechpro.com

Михаил Борисович Столбов — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ-инновации“, Санкт-Петербург; старший научный сотрудник; E-mail: stolbov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 656.25-52:656.22.05

С. В. Бибиков, Ю. Н. Матвеев, Н. Н. Семенов

ОЦЕНКА ФУНКЦИОНАЛЬНОЙ БЕЗОПАСНОСТИ ОБНАРУЖЕНИЯ ВИБРОАКУСТИЧЕСКОГО СИГНАЛА ПРИБЛИЖАЮЩЕГОСЯ ПОЕЗДА

Исследована функциональная безопасность обнаружения виброакустического сигнала приближающегося поезда методом энергетического обнаружителя. Найдено нижнее значение порога обнаружения исходя из предложенной вероятности ложной тревоги, доказана достаточность применения разработанного метода обнаружения приближающегося поезда для бесстыковых путей.

Ключевые слова: *приближающийся поезд, виброакустический сигнал, энергетический обнаружитель.*

Введение. Для того чтобы устройство, обнаруживающее по специфическим виброакустическим колебаниям рельса приближающийся поезд, могло применяться в ОАО „РЖД“ при подаче сигналов оповещения, оно должно удовлетворять достаточно жестким требованиям по надежности и функциональной безопасности. В качестве основного критерия функциональной безопасности используется среднее время наработки на опасный отказ: не менее 10^6 ч. Критериями опасного отказа являются отсутствие включения сигнала оповещения до начала установленного времени упреждения сигнализации (от включения сигнала оповещения до достижения первой осью приближающегося поезда места установки устройства оповещения) или неразборчивый сигнал. Установленное время упреждения сигнализации равно 50 с.

Среднее время наработки на опасный отказ определяется тремя показателями:

1) вероятность пропуска обнаруживаемого приближающегося поезда при наличии помех;

2) вероятность опасного необнаруживаемого или невосстанавливаемого отказа аппаратной реализации (10^{-11}—10^{-12} 1/с [1]);

3) вероятность восстанавливаемого опасного отказа („сбоя“) с нерегламентированным временем восстановления с учетом обнаружения и восстановления работоспособности средствами операционной системы реального времени.

Поскольку надежность системы оповещения обеспечивается средствами дублирования аппаратуры, показатель *3* можно не оценивать отдельно от показателя *2*.

В настоящей статье рассматривается только ситуация пропуска обнаруживаемого приближающегося поезда при наличии помех, так как она является решающей при доказательстве безопасности устройства оповещения. Вследствие случайного характера помех принципиально

невозможно добиться их полного устранения. Путем совершенствования приемных устройств можно снизить вероятность ошибки обнаружения полезного сигнала только до некоторого приемлемого уровня [2].

Классический энергетический обнаружитель. Постановка задачи. Воспользуемся классической постановкой задачи обнаружения сигнала по уровню его энергии U. Пусть на выходе одного частотного канала имеется некий сигнал — случайный процесс:

$$U(t)=V(t)+z(t),$$

он может представлять либо только помеху $z(t)$, либо сумму сигнала $V(t)$ и помехи. Будем считать, что наличие сигнала $V(t)$ тоже случайно.

Для решения вопроса о наличии сигнала в данный момент времени можно принять: сигнал присутствует, если $U(t)>E$, т.е. превышает некоторое пороговое значение, в противоположном случае сигнал отсутствует.

Ошибочный ответ может быть получен:

1) когда сигнал отсутствует, $V(t) = 0$, но уровень помехи превышает уровень E (событие А, „ложная тревога“);

2) когда сигнал присутствует, $V(t) \neq 0$, но сумма сигнала и помехи не превышает уровня $U(t)$ (событие В, „пропуск сигнала“).

Вероятность ложной тревоги, т.е. того, что будут совмещены отсутствие сигнала и превышение помехой уровня E при отсутствии сигнала, равна априорной вероятности q отсутствия сигнала, умноженной на апостериорную вероятность P превышения уровня E, при условии, что сигнал отсутствует. Зададим значение q, а значение P легко получить по одномерной функции распределения помехи $W(x)$:

$$P\left(U > E\right) = \int_{E}^{\infty} W(x)dx, \tag{1}$$

тогда

$$P_{\text{A}} = q\int_{E}^{\infty} W(x)dx.$$

Вероятность того, что будут совмещены присутствие сигнала и непревышение уровня E суммарным напряжением (вероятность события В), равна априорной вероятности присутствия сигнала, умноженной на апостериорную вероятность непревышения уровня E при условии, что сигнал присутствует.

Априорная вероятность присутствия сигнала равна:

$$p=1-q.$$

Апостериорную вероятность непревышения уровня E можно получить, используя одномерную функцию распределения суммы сигнала и помехи $W_1(x,V)$.

$$P\left(U \le E\right) = \int_{-\infty}^{E} W_1(x,V)dx,$$

тогда

$$P_{\text{B}} = p\int_{-\infty}^{E} W_1(x,V)dx.$$

Так как события А и В несовместны, то вероятность получения ошибочного ответа P_{A} или P_{B}) равна

$$P_{\text{А\_или\_В}} = P_{\text{А}} + P_{\text{В}} =$$

$$= q\int_{E}^{\infty} W(x)dx + p\int_{-\infty}^{E} W_1(V,x)dx = 1 - \left[p\int_{E}^{\infty} W_1(V,x)dx + q\int_{-\infty}^{E} W(x)dx \right].$$

Искомая вероятность получения правильного ответа P_0 равна:

$$P_0(E,V) = 1 - P_{\text{А\_или\_В}} = p\int_{E}^{\infty} W_1(x,V)dx + q\int_{-\infty}^{E} W(x)dx. \qquad (2)$$

Рассмотрим задачу нахождения оптимальной величины порога E, для которого вероятность получения правильного ответа (2) при заданных функциях распределения сигнала и помехи максимальна. Вычислив производную выражения (2) по E и приравняв ее нулю, получим уравнение для определения оптимального уровня:

$$\frac{dP(E)}{dE} = 0, \qquad (3)$$

откуда $qW(E) = pW_1(E,V)$.

Статистический критерий, обеспечивающий максимальную вероятность получения правильного ответа при одном или нескольких измерениях, называется критерием „идеального наблюдателя“. Рассмотрим решение уравнения (3) на примере обнаружения положительной телеграфной посылки (положительного импульса с амплитудой V) на фоне помехи, подчиняющегося нормальному закону распределения, с дисперсией σ^2. Наличие или отсутствие сигнала скажется только на среднем значении суммарного сигнала.

Соответственно плотности распределения вероятности будут иметь вид:

$$W(U) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{U^2}{2\sigma^2}\right),$$

$$W_1(U,V) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(U-V)^2}{2\sigma^2}\right).$$

В случае, когда априорная вероятность появления сигнала неизвестна, часто полагают p=1/2, считая, что равновероятно как наличие, так и отсутствие сигнала. Заметим, что при этом q=1/2, тогда для описанных распределений $E = V/2$.

Если уровень E выбран, то для вероятностей ложной тревоги и пропуска сигнала, используя приведенные формулы, получим выражения:

$$P_{\text{А}} = q\left[1 - \Phi\left(\frac{E}{\sigma}\right)\right], \quad P_{\text{В}} = p\left[1 - \Phi\left(\frac{U-E}{\sigma}\right)\right], \qquad (3)$$

где $\Phi(y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp\left(-\frac{x^2}{2}\right)dx$ — функция Крампа.

В радиолокационных системах во избежание ложного обнаружения цели используется критерий Неймана—Пирсона. Обычно задают значение вероятности ложной тревоги порядка 10^{-5}. Эта величина не регламентирована руководящими документами по железнодорожной безопасности, но используется в экспертной практике. При использовании критерия Неймана—Пирсона значение вероятности ложной тревоги фиксируется изначально. Так как эта вероятность функционально связана с относительным порогом, то последний также оказывается заданным.

Такой подход позволяет удовлетворить одновременно двум противоречивым требованиям:

1) чтобы вероятность пропуска сигнала не превосходила некоторой заданной величины,

2) чтобы вероятность ложной тревоги была минимальной.

Логика такого выбора представлена на рис. 1.

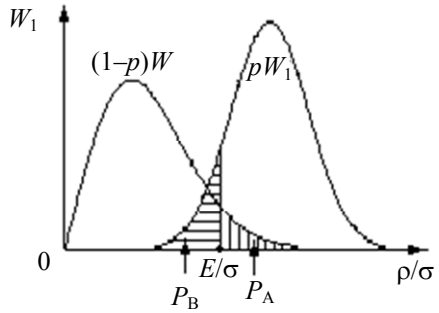


Рис. 1

На рис. 2, *а* приведена зависимость вероятности ложной тревоги от дисперсии шума случайного процесса σ при отсутствии сигнала поезда. Из рисунка видно, что для достижения $P_A=10^{-5}$ достаточно установить порог обнаружения 3σ.

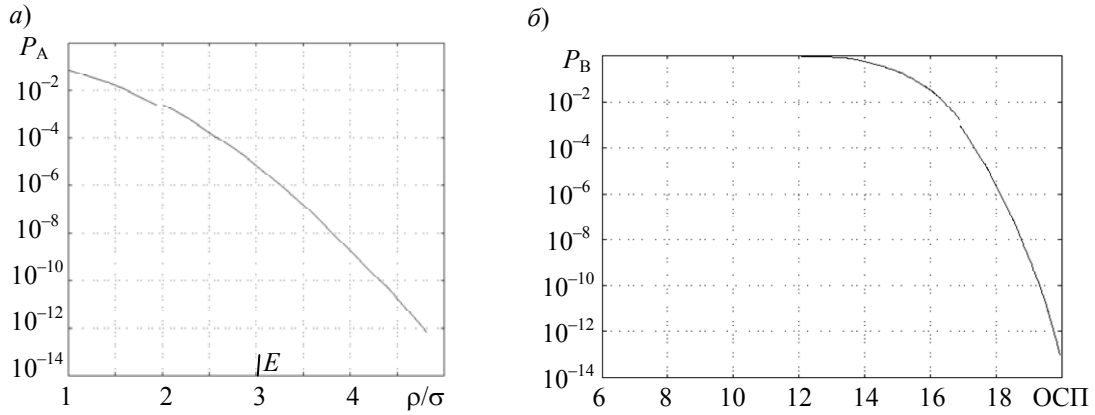


Рис. 2

Рассчитанная зависимость вероятности пропуска цели от отношения сигнал—помеха (ОСП) приведена на рис. 2, *б*. Видно, что при ОСП=19 дБ вероятность пропуска цели составляет $<10^{-8}$, это значение резко снижается при увеличении ОСП. При ОСП=20 дБ и более вероятность пропуска цели составляет существенно меньшие значения, чем необходимо для безопасного функционирования устройства.

Исследования экспериментальных данных. На рис. 3, *а* приведена спектрограмма исходного сигнала приближающегося со скоростью 60 км/ч поезда. Фонограмма на бесстыковом пути без неоднородностей с железобетонными шпалами получена устройством оповещения типа „Сигнализатор П“. Запись велась на участке пути с высоким уровнем шума окружающей среды. Проезд проехал место установки устройства 160-й секунде. Из рисунка видно, что до частоты 11 кГц наблюдаются тональные составляющие, являющиеся гармониками наведенных электромагнитных сигналов [3]. Основная часть энергии сигнала поезда расположена в полосе 12 000—19 400 Гц. Резкий рост спектральных компонент на 83-й секунде связан со ступенчатым возрастанием амплитуды входного сигнала на 4—5 дБ при проезде изолирующего стыка и уравнительных пролетов и въезде на блок — участок пути, на котором установлено устройство оповещения. На 140-й секунде происходит насыщение усилительных каскадов устройства, затем — насыщение внутреннего усилителя датчика, в результате чего происходит перераспределение энергии по всему спектру.

Чтобы убрать ненужные и мешающие обнаружению сигналы, проведем полосовую фильтрацию исходного сигнала в диапазоне 11 000—19 400 Гц (рис. 3, *б*).

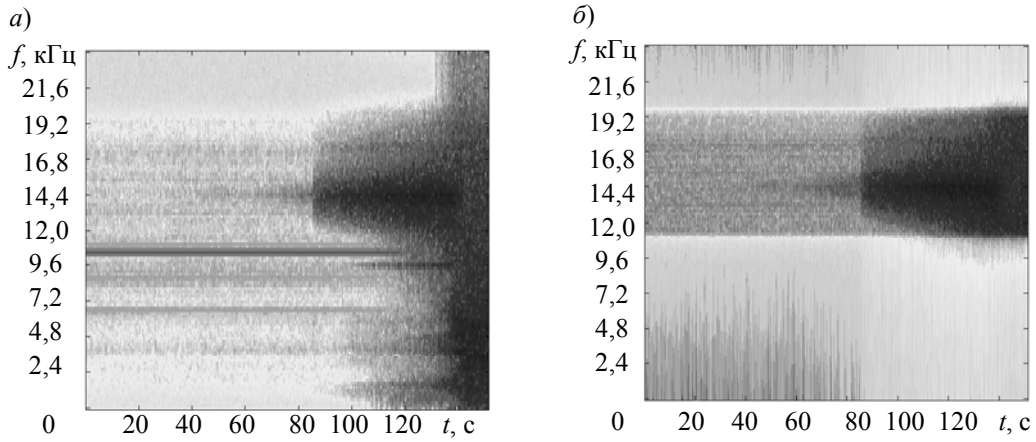


Рис. 3

Полученный полосовой сигнал в начале записи содержит равномерный по спектру шум, затем на фоне этого шума появляется сигнал движущегося поезда.

Рассмотрим процедуру оценки требуемого значения вероятности пропуска цели. Время наработки на опасный отказ, согласно требованиям безопасности, составляет $T_{оо} = 10^6$ ч = $= 3{,}6 \cdot 10^9$ с.

Поскольку „Сигнализатор П“ является изделием периодического использования и не может принципиально включать сигнал о приближении поезда ежесекундно, оценим предельную периодичность включения сигнала оповещения: $T_{оп} = T_{ср} \cdot 2 + T_{обн} + T_{пр} = 50 \cdot 2 + 5 \cdot 2 + 1 = 111$ с. Здесь $T_{ср}$ — время от срабатывания устройства до проезда поезда по месту установки, в это время оно подает сигналы оповещения и не может обнаруживать другой поезд; $T_{обн}$ — время работы алгоритма (время обнаружения и принятия решения), оно зафиксировано на уровне 5 с. Процесс выключения сигналов оповещения считаем симметричным процессу включения; $T_{пр}$ — время проезда (берется время проезда локомотива на предельной скорости ≈1 с).

Рассматривается случай непрерывного (один за одним) движения поездов. В реальности применение изделия и работа на путях перестают быть эффективными при интервале следования поездов менее 10 мин. Коэффициент снижения интенсивности отказов из-за периодичности и уменьшения времени наблюдения: $K_{оо} = T_{оп} T_{нб} = 111$, где $T_{нб} = 1$ с — время наблюдения.

Оценим вероятность опасного отказа с учетом периодичности работы изделия. Интенсивность отказов определяется как:

$$\lambda_{оо} = K_{оо} \cdot 1/T_{оо} = 111/3{,}6 \cdot 10^9 = 3{,}08 \cdot 10^{-8} \text{ 1/с}.$$

Учитывая малость $\lambda_{от}$, вычислим требуемое значение вероятности пропуска цели, которое ограничит помехоустойчивость метода в отношении пропуска сигнала приближения поезда.

$$P_{оо} = \lambda_{оо} T_{нб} = 3{,}08 \cdot 10^{-8} \text{ 1/с}.$$

Согласно рис. 3, момент принятия решения для получения данного значения вероятности ОСП=19 дБ.

На рис. 4 приведена зависимость ОСП сигнала приближения поезда, отфильтрованного полосовым фильтром, от времени (см. рис. 3, *б*).

Значение энергии помехи вычислялось в паузе перед приближением поезда. Включение оповещения по порогу 3σ происходит на 78-й секунде. Время сигнализации составило 82 с. Из рис. 4 видно, что даже при высокой зашумленности выбранного участка эксперимента и при фиксированной широкополосной фильтрации значение ОСП на 110-й секунде намного больше безопасного порогового значения 19 дБ, обеспечивающего требуемые вероятность пропуска цели и среднее время наработки на опасный отказ.

Среднеквадратичное отклонение времени сигнализации для различных типов поездов в различных погодных условиях (но без снега), при установке устройства оповещения на одном и том же месте на рельсе, составило 4,2 с при общем количестве поездов 34.

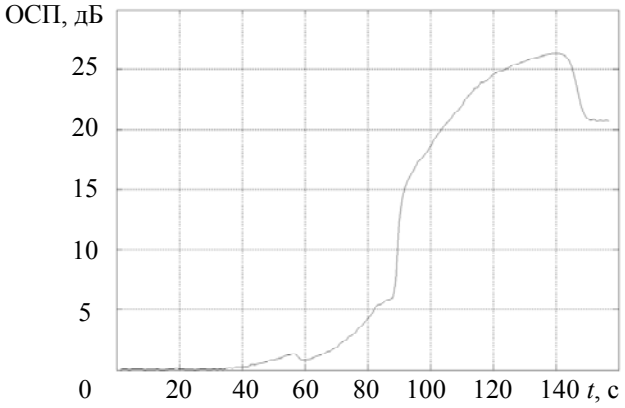


Рис. 4

Заключение. Для бесстыкового пути без неоднородностей и мешающих факторов использования классического энергетического обнаружителя достаточно как для обеспечения требуемого времени сигнализации при обнаружении виброакустического сигнала приближающегося поезда, так и для доказательства надежности алгоритма обнаружения приближающегося поезда по критерию вероятности пропуска цели.

СПИСОК ЛИТЕРАТУРЫ

1. *Богатырев В. А., Бибиков С. В.* Оценка функциональной безопасности дублированных вычислительных систем // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 2(78). С. 146.

2. *Бакут П. А., Большаков И. А., Герасимов Б. М.* и др. Вопросы статистической теории радиолокации / Под общ. ред. *Г. П. Тартаковского*. М.: Сов. радио, 1964. Т. 1. 426 с.

3. *Бибиков С. В., Шапарь А. В.* Движущийся поезд как источник звуковых волн, распространяющихся по рельсовому пути // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 5(81). С. 152.

Сведения об авторах

Сергей Викторович Бибиков — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ“, Санкт-Петербург; заместитель директора департамента; E-mail: bibikov@speechpro.com,

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

Николай Николаевич Семенов — канд. техн. наук, доцент; ООО „ЦРТ“, Санкт-Петербург; руководитель группы; E-mail: semenov-n@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 621.391.037.372

М. Б. Столбов, М. Ю. Татарникова

РАЗДЕЛЕНИЕ РЕЧИ ЦЕЛЕВОГО И СТОРОННИХ ДИКТОРОВ С ИСПОЛЬЗОВАНИЕМ ДВУХМИКРОФОННОЙ СИСТЕМЫ

Рассмотрен метод разделения речи целевого и сторонних дикторов с помощью обработки сигналов двух симметричных микрофонов, ориентированных в противоположных направлениях. Алгоритм обработки обеспечивает пространственное разделение дикторов.

Ключевые слова: *детектор речи, двухканальная фильтрация аудиосигналов.*

Введение. Одной из актуальных практических задач в области обработки речевых сигналов является контроль соблюдения речевого регламента сотрудниками call-центров, диспетчерами (операторами). Данная задача может решаться методами автоматического распознавания речи. На достоверность результатов существенное влияние оказывает окружающая акустическая обстановка, в частности, речь сторонних дикторов, которая может интерпретироваться как речь целевого диктора (РЦД). Для улучшения результатов распознавания необходимо применять специальные методы выделения РЦД.

Для выделения РЦД в сложной акустической обстановке применяются различные системы, включающие в себя как аппаратную, так и алгоритмическую составляющие.

Традиционно для выделения РЦД используются индивидуальные гарнитуры и микрофоны. При этом уровень речи целевого диктора настолько превосходит уровень окружающих помех, что дополнительной обработки сигналов не требуется. Однако в ряде случаев из-за условий работы операторов такой подход не может быть использован.

Также используются микрофонные решетки и алгоритмы пространственной фильтрации [1]. Однако эти методы непригодны для больших помещений и требуют чрезмерных аппаратных затрат.

В последнее время для мониторинга совещаний все большее применение находят распределенные системы микрофонов (claster of microphones), когда каждый диктор снабжается индивидуальным (close-talk) микрофоном [2—5]. Поскольку в таких системах в каждый из индивидуальных микрофонов попадает речь сторонних дикторов, выделение РЦД осуществляется на основе совместной обработки сигналов всех микрофонов. Такие системы требуют оборудования микрофонами всех рабочих мест операторов, при этом предполагается, что сторонние дикторы (операторы) всегда находятся в непосредственной близости к «индивидуальным» микрофонам. Эти ограничения соответствуют сценарию совещания и задаче распознавания речи дикторов (meeting speech recognition) [6], однако не подходят для ситуаций, когда сторонние дикторы меняют местоположение.

Перед авторами стояла задача создания гибкой системы, обеспечивающей выделение РЦД на оборудованных рабочих местах без использования совместной обработки сигналов всех микрофонов. Основная идея предложенного подхода заключается в использовании в рабочей зоне оператора микрофонного блока, состоящего из двух симметричных микрофонов, один из которых направлен в сторону оператора и предназначен для записи его речи, а другой направлен в противоположную сторону и предназначен для приема посторонних звуков.

В работе описан алгоритм совместной обработки сигналов микрофонов, позволяющий выделить речь целевого диктора на фоне акустических шумов окружения и речи сторонних дикторов.

Микрофонный блок. Для выделения речи оператора была предложена схема с двумя противоположно направленными микрофонами. Микрофон основного канала *(m)* предназначался для получения аудиосигнала от оператора, микрофон опорного канала (*r*) — для получения сигнала от акустического окружения. Для избежания проблем калибровки (выравнивания АЧХ микрофонов) и упрощения алгоритмов обработки были использованы симметричные (одинаковые) микрофоны (рис. 1). Для увеличения эффективности пространственного выделения речи оператора и снижения влияния шумов окружения были выбраны суперкардиоидные микрофоны.

а) *б*)

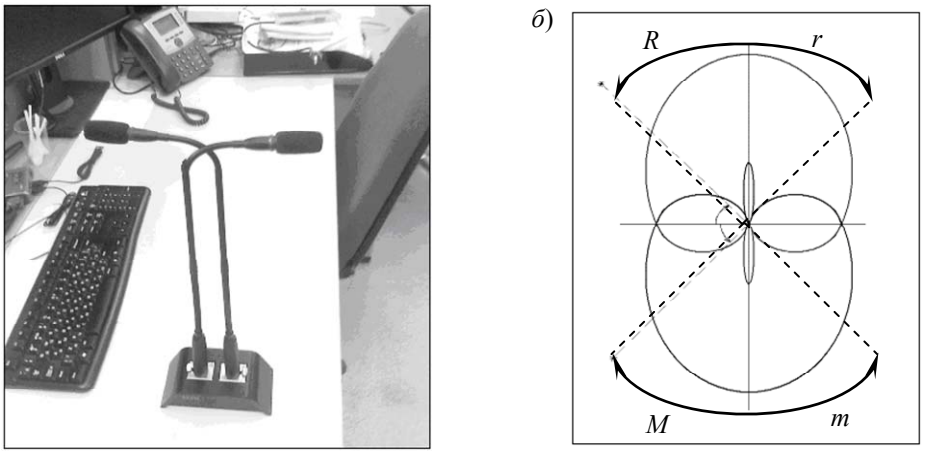


Рис. 1

Выделение речи целевого диктора. Предполагалось, что основным источником помехи является речь стороннего диктора.

Рассматривались два варианта обработки аудиофайла: фильтрация сигналов с целью компенсации речи сторонних дикторов [5, 7] и детектирование участков сторонних дикторов (overlapped speech detection) с целью их исключения из процесса распознавания речи [4—6, 8—11].

Поскольку фильтрация обычно приводит к искажениям РЦД, нами был выбран метод детектирования. Пригодными для дальнейшей обработки и распознавания речи считались сегменты сигнала, содержащие только речь целевого диктора.

Алгоритм выделения РЦД работает следующим образом. На участках речи оператора, не содержащих помех, исходный сигнал основного канала сохраняется, на остальных участках сигнал подавляется на заданную величину (обычно 20 дБ).

Для каждого кадра сигнала микрофона основного канала принималась одна из следующих гипотез:

— H0: паузы речи,

— H1: РЦД,

— H2: речь стороннего диктора,

— H12: одновременная речь целевого и стороннего дикторов.

В работе [11] описан алгоритм сегментации сигнала согласно этим гипотезам на основе оценок критерия отношения мощностей сигналов микрофонов. Предложенный алгоритм обеспечивает пространственную фильтрацию сигналов РЦД и стороннего диктора.

Пример распределения мощностей сигналов основного и опорного каналов (P_r и P_m) для разных участков сигнала приведен на рис. 2 (квадраты — РЦД, кружки и крестики — речь сторонних дикторов).

Эксперименты на записях, выполненных в лабораторных условиях, подтвердили эффективность метода детектирования и алгоритма обработки. Однако в реальных условиях

опытная эксплуатация разработанной системы обнаружила слабые места предложенного подхода.

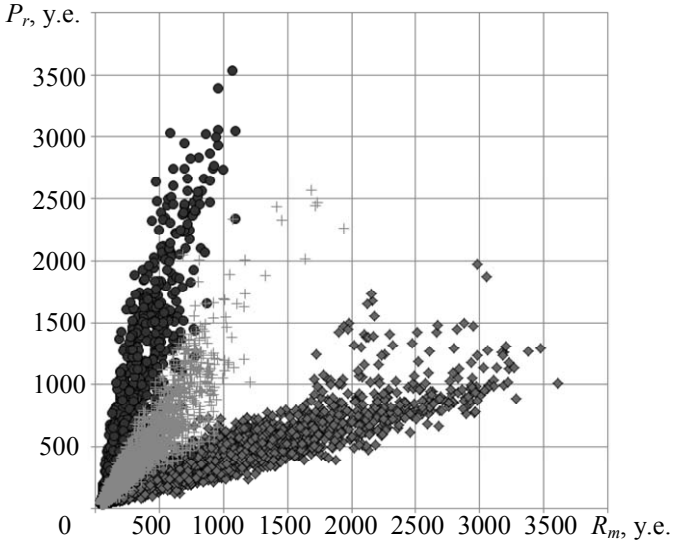


Рис. 2

Выделение РЦД на фоне шумов окружения. Одной из основных проблем выделения РЦД в реальных условиях оказалось большое число источников акустических сигналов вокруг оператора, таких как шум компьютеров, шум принтера, речь сторонних операторов, звук клавиатуры операторов, шелест бумаг, работающие аудиоустройства (динамики, телевизор), наводки и звонки мобильных телефонов и др.

Предложенный ранее критерий, основанный на оценке мощностей основного и опорного каналов, идентифицировал такие звуки как РЦД. Кроме того, речевые сигналы удаленных дикторов поступают в основной и опорный микрофоны приблизительно в равных пропорциях, поэтому критерий отношения мощностей, предложенный в работе [11], оказался неэффективным.

Задача заключалась в выборе новых критериев, позволяющих идентифицировать РЦД в сложной акустической обстановке. Для детектирования участков РЦД на фоне неречевых помех могут быть использованы одноканальные критерии (тональность сигналов, частота пересечения нуля и др.). Однако эти критерии не позволяют выделить РЦД на фоне речевых сигналов (речь удаленных дикторов, речевые сигналы аудиоустройств).

Для выявления акустических событий, порождаемых удаленными источниками, целесообразно использовать так называемые кроссканальные критерии (cross-channel features).

В работах, посвященных детектированию РЦД на фоне речи сторонних дикторов, предложены несколько групп критериев отбора:

— на основе функции когерентности [9],

— на основе функций кросскорреляции [4, 6, 10],

— на основе функций кросс-спектра [8],

— на основе мер подобия амплитудных спектров и спектров мощности [3, 5].

Сегментация сигнала на отдельных кадрах осуществляется по превышению порога одним или несколькими критериями.

Критерии на основе кросс-спектров. По результатам предварительных исследований групп критериев для системы двух микрофонов нами были предложены модифицированные критерии на основе кросс-спектров.

Введем обозначения оценок следующих функций: $\Phi_{mr}(k, t)$ — оценка мгновенного кросс-спектра сигналов основного и опорного каналов, $\langle\Phi_{mr}(k, t)\rangle$ — усредненная по времени оценка кросс-спектра, $\langle\Phi_{mm}(k, t)\rangle$, $\langle\Phi_{rr}(k, t)\rangle$ — усредненная по времени оценка спектров мощности сигналов основного и опорного каналов, k — индекс частоты, t — индекс кадра данных.

На основе этих функций определим интегральные критерии:

$$Q_1 r(t) = \frac{1}{|F|} \sum_k \frac{|< \Phi_{mr}(k,t) >|}{< \Phi_{mm}(k,t) >},$$

$$Q_1 m(t) = \frac{1}{|F|} \sum_k \frac{|< \Phi_{mr}(k,t) >|}{< \Phi_{rr}(k,t) >}.$$

Проверка кросс-спектральных критериев на ряде фонограмм, сделанных в реальных условиях, показала возможность их использования для разделения близких и удаленных источников акустических сигналов. Пример для критериев $Q_1m(t)$, $Q_1r(t)$ приведен на рис. 3.

Видно, что критерий $Q_1m(t)$ позволяет детектировать РЦД (участки 1—3,5; 12—14,5; 20—22 с), $Q_1r(t)$ — речь находящегося вблизи стороннего диктора (6—11,5; 15—15,5 с) в присутствии фонового звука телевизора. На участке звука телевизора в отсутствие речи (участок 3,5—5,5 с) оба критерия оказались нечувствительными к звуку телевизора.

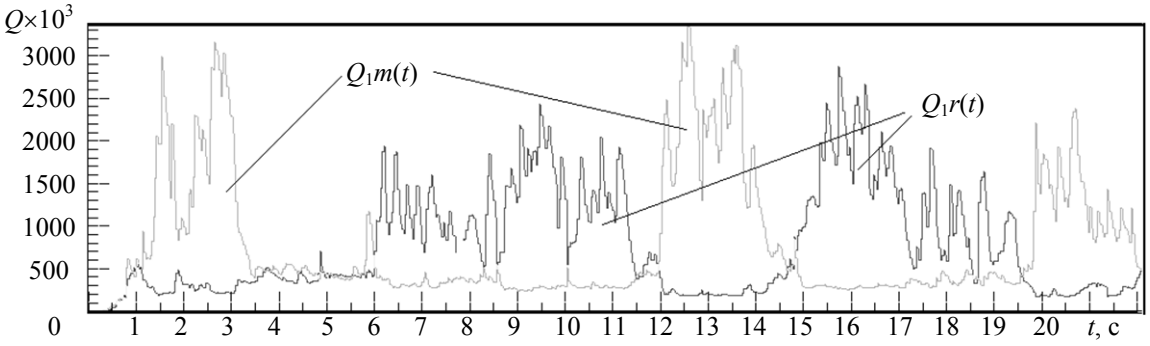


Рис. 3

Исследования на реальных записях показали робастность кросс-спектральных критериев по отношению к акустическим сигналам удаленных источников окружения. Исследование области применения этой группы критериев является предметом дальнейшей работы.

Заключение. Предложен метод выделения участков речи целевого диктора на фоне речи сторонних дикторов и акустических помех. Основная идея предложенного метода заключается в детектировании участков сигнала с речью целевого диктора и исключении участков помех и речи сторонних дикторов.

Предложена система из двух симметричных противоположно направленных суперкардиоидных микрофонов. Алгоритм детектирования речи целевого диктора основан на оценке кроссканальных критериев.

Предложен кросс-спектральный критерий, позволяющий разделить близко расположенные и удаленные источники речевых сигналов. Основным достоинством предложенной системы является простота и применимость в широком диапазоне практических ситуаций, ограничением (по сравнению с использованием гарнитуры) — потеря фрагментов речи целевого диктора на участках присутствия акустических помех и речи близко расположенного стороннего диктора. Более полное исследование предложенного критерия, в частности его применимость для выделения РЦД в сложной акустической обстановке при совместном использовании нескольких критериев, является предметом дальнейшей работы.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Morgan P., George E., Lee T., Kay M.* Co-Channel Speaker Separation // Proc. of the ICASSP. 1995. Vol. 1. P. 828—831.

2. *Nasu Y., Shinoda K., Furui S.* Cross-Channel Spectral Subtraction for meeting speech recognition // Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2011. P. 4812—4815.

3. *Xiao B.* et al. Overlapped speech detection using long-term spectro-temporal similarity in stereo-recording // Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2011. P. 5216—5219.

4. *Kumatani K.* et al. Channel selection based on multichannel cross-correlation coefficients for distant speech recognition // Proc. of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2011). UK, 2011. P. 1—6.

5. *Yakoyama R.* et al. Overlapped Speech Detection in Meeting Using Cross-Channel Spectral Subtraction and Spectrum Similarity // Proc. Interspeech. 2012. P. 1—4.

6. *Boakye K., Stolcke A.* Improved Speech Activity Detection Using Cross-Channel Features for Recognition of Multiparty Meetings // Proc. Interspeech. USA, 2006. P. 1962—1965.

7. *Cao Y., Sridman S., Moody M.* Multichannel Speech Separation by Eigendecomposition and Its Application to Co-Talker Interference Removal // IEEE Trans. on SAP. 1997. Vol. 5, N 3. P. 209—219.

8. *Wrigley S. N., Brown J., Wan V., Renals S.* Speech and Crosstalk Detection in Multichannel Audio // IEEE Trans. on SAP. 2005. Vol. 13, N 1. P. 84—91.

9. *Yen K.-C., Zhao Y.* Robust Automatic Speech Recognition using a multi-channel signal separation front-end // Proc. Fourth Intern. Conf. on Spoken Language, ICSLP. 1996. Vol. 3. P. 1337—1340.

10. *Laskowski K., Schulttz T.* A geometric interpretation of non-target-normalized maximum cross-channel correlation for vocal activity detection in meetings // Proc. of NAACL HLT. 2007. P. 89—92.

11. *Stolbov M., Tatarnikova M.* Speech and Crosstalk Detection for Robust Speech Recognition Using a Dual Microphone System // Proc. of 15th Intern. Conf. on Speech and Computer, SPECOM. 2013. P. 310—318.

Сведения об авторах

Михаил Борисович Столбов — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ-инновации“, Санкт-Петербург; старший научный сотрудник; E-mail: stolbov@speechpro.com

Марина Юрьевна Татарникова — ООО „ЦРТ-инновации“, Санкт-Петербург; старший научный сотрудник; E-mail: tatmar@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

СИСТЕМЫ РАСПОЗНАВАНИЯ ЛИЧНОСТЕЙ ПО ГОЛОСУ

УДК 343.98

Е. В. Булгакова, Е. В. Краснова

ЭКСПЕРТНЫЕ СИСТЕМЫ И МЕТОДЫ ИДЕНТИФИКАЦИИ ДИКТОРА

Проанализированы методы проведения фонографической экспертизы. Рассмотрено используемое в целях идентификации диктора программное обеспечение. Описывается звуковой редактор SIS II.

Ключевые слова: *криминалистическая фонографическая экспертиза, идентификация дикторов, программное обеспечение.*

Введение. В мировой криминалистической практике в настоящее время фонографическая экспертиза, несомненно, играет важную роль. Однако в России и за рубежом наблюдается различие в подходах к проведению данного вида криминалистической экспертизы, а также в технических средствах для идентификации диктора. Традиция использования различных технических средств и подходы к разработке программного обеспечения также различаются. В отличие от отечественной традиции, за рубежом при проведении фонографической экспертизы решающую роль играют личность эксперта и степень доверия к его профессиональной квалификации. В связи с этим за рубежом эксперты, как правило, применяют программное обеспечение, не предназначенное непосредственно для задачи идентификации диктора. Обычно в качестве таких программ выступают звуковые редакторы с относительно стандартным функционалом. Также некоторыми зарубежными экспертами предлагается использовать при проведении фонографической экспертизы программное обеспечение, разработанное для специалистов в других научных областях (например, в области медицины). В России специально для криминалистических целей разработан ряд программных комплексов, широко используемых правоохранительными органами. Однако функционал этих программ также варьирует, предоставляя экспертам различный набор исследовательских возможностей при проведении фонографической экспертизы [1].

В настоящей работе анализируются методы идентификации диктора, принятые в криминалистической практике, рассматриваются преимущества использования звукового редактора SIS II, разработанного в ООО „ЦРТ“.

Под автоматическим подходом как в отечественной, так и в зарубежной традиции понимается процесс сравнения фонограмм и принятия решения о тождестве или различии звучащих на них голосов без участия эксперта [4].

Аудитивный метод идентификации диктора в отечественной традиции подразумевает, в первую очередь, работу эксперта-фонетиста. За рубежом под аудитивным методом идентификации понимается любая идентификация, не привлекающая дополнительных инструментов. Таким образом, в понятие „аудитивный метод“ в зарубежной традиции входит как прослуши-

вание опытным фонетистом предоставленных спорных фонограмм и их сравнение с фонограммами речи подозреваемых, так и прослушивание фонограмм „наивным“ слушателем.

Акустический подход в отечественной традиции и акустико-фонетический в зарубежной также могут быть соотнесены. Однако в отечественной традиции акустико-фонетический анализ сегментного и суперсегментного уровней может использоваться в рамках лингвистического подхода, в то время как акустический подход в зарубежной традиции предполагает только инструментальный анализ физических параметров звукового сигнала без непосредственной соотнесенности с лингвистической информацией.

Лингвистический подход не получил за рубежом отражения в том виде, в котором он широко применяется в отечественной экспертной практике (с использованием признакового пространства). И, наконец, зарубежный спектрографический метод в отечественной традиции развит не был.

Анализ речевого сигнала с использованием различных технических средств как в России, так и за рубежом, проводится в рамках различных подходов.

В зарубежной криминалистической практике программы, используемые для решения задач идентификации диктора, — это, как правило, звуковые редакторы [5], функционал которых включает возможность визуализации речи (например, Praat, Multi-speech и Wave Surfer). Программа Praat, разработанная в Амстердамском университете, создавалась в первую очередь для научно-исследовательских целей, она предназначена для лингвистов, исследующих звучащую речь. Программа предоставляет возможности для визуализации речевого сигнала, сегментации речевого потока, анализа и синтеза речи. Программа Multi-speech, разработанная компанией KayPentax, — это компьютерная речевая лаборатория, в которой имеются возможности анализа речи, инструменты для визуализации речевого сигнала различными способами. Она предназначена для акустического анализа речевого сигнала, а также анализа устной речи в медицинских целях. Звуковой редактор Wave Surfer, разработанный в Королевском технологическом институте (Стокгольм), активно используется для решения задач, связанных с акустической фонетикой, благодаря широким возможностям визуализации речи, а также возможности считывать и записывать различные форматы файлов транскрипции.

Некоторые зарубежные специалисты в области идентификации диктора предлагают использовать для решения идентификационных задач программы, изначально предназначенные для иных целей. Например, программа Glottex, первоначально разработанная для использования в медицине, позволяет получать информацию о свойствах источника голоса, включая физическую структуру голосовых складок. По мнению авторов статьи [6], эти данные могут быть использованы для судебного сравнения голосов.

Однако ошибочно было бы утверждать, что за рубежом не существует программных комплексов, разработанных специально для целей фонографической экспертизы. В качестве примера таких разработок можно привести программу SAUSI (Semi-Automatic Speaker Identification System), созданную специалистом в области идентификации диктора Г. Холлиеном. В программе анализируются составляемые на основе фонограмм векторы значений, включающие четыре модификации: спектр длительных фрагментов, частота основного тона, временное распределение энергии, характерные особенности формант гласных. На первом этапе четыре приведенных вектора подвергаются самостоятельному анализу, а затем — в сочетаниях друг с другом, на основе чего формируются результаты сравнительного анализа [7].

Одной из систем разработанных непосредственно для идентификации диктора при проведении фонографической экспертизы, является система OTExpert [8], разработанная специалистами ООО „ОТ-Контакт“ (Москва) и используемая в некоторых криминалистических лабораториях Министерства юстиции РФ. Система предназначена для инструментального криминалистического исследования фонограмм. Функционал программы включает возможности визуализации (осциллограмма, спектрограмма, кепстрограмма, КЛП-спектрограмма с прорисовкой

локальных максимумов), сегментации и редактирования звукового сигнала, автоматического статистического анализа параметров основного тона и формант, а также используемого в работе эксперта сравнительного анализа речевых сигналов на разных уровнях.

Разработанный специалистами ООО „Целевые технологии“ [9] комплекс Justiphone представляет собой систему криминалистического исследования фонограмм речи, обеспечивающую шумоочистку фонограмм, установление дословного содержания, выявление идентификационных признаков говорящего, техническое исследование фонограмм на предмет наличия или отсутствия признаков монтажа. Программный комплекс имеет гибкую конфигурацию. Наиболее интересной с точки зрения идентификации диктора является комплектация Justiphone-voice analyzer, которая обеспечивает исследование акустических параметров голоса и звучащей речи. Justiphone-voice analyzer рассчитывает статистические характеристики основного тона и оценивает средневзвешенное относительное отклонение полученных параметров спорной фонограммы от параметров фонограммы-образца.

Автоматизированная система „Диалект“ основана на методике идентификации диктора, разработанной ведущими отечественными специалистами в области фонографической экспертизы [10, 11], включающей акустический и лингвистический методы исследования фонограмм (аудитивный анализ в данном случае является составляющей лингвистического метода). Функционал программы позволяет эксперту обрабатывать речевые сигналы, вычислять акустические и выявлять лингвистические признаки, сравнивать параметры и принимать решения по проводимому исследованию. Эксперт имеет возможность проводить углубленный анализ микроструктуры речевого сигнала и акустических шумов фонограммы на этапах определения пригодности фонограмм для идентификационных исследований и идентификации устной речи неизвестного и подозреваемого.

Итак, как показал анализ, при проведении фонографической экспертизы используется широкий спектр различных программных средств. Однако в описанных выше программах не предусмотрена возможность автоматического принятия идентификационного решения. Эта функция реализована в специализированном звуковом редакторе SIS II v2.0 (далее SIS II), разработанном ООО „ЦРТ“. Наличие модуля обобщенного решения отличает этот редактор от других программ, предназначенных для проведения идентификации диктора.

Редактор SIS II предназначен для анализа речевых сигналов, их шумоочистки и автоматизации выполнения криминалистических экспертиз фонограмм на всех этапах. SIS II обладает уникальными инструментами для идентификационного исследования. Биометрические алгоритмы и экспертные модули позволяют автоматизировать и формализовать многие процессы фонографического исследования, например поиск одинаковых слов и звуков, отбор сравниваемых звуковых и мелодических фрагментов, сравнение дикторов по формантам и основному тону, аудитивный и лингвистический типы анализа. Результаты по каждому методу исследования представляются в виде численных показателей общего идентификационного решения. Рассмотрим модули редактора.

1. Модуль автоматической идентификации (рис. 1) позволяет выполнять сравнение в режиме „один-к-одному“ с использованием следующих алгоритмов: спектрально-формантного, статистики основного тона, смеси гауссовых распределений, обобщенного метода.

Значения вероятности совпадения и различия дикторов рассчитываются не только для каждого из методов, но и для их совокупности. Все результаты сравнения речевых сигналов, получаемые в модуле автоматической идентификации, основаны на выделении в них идентификационно значимых признаков и вычислении меры близости между полученными наборами признаков.

2. Модуль сравнения формант. Процесс идентификации с использованием модуля (рис. 2) может быть разделен на два этапа. Сначала эксперт осуществляет поиск и отбор опорных звуковых фрагментов:

— по диаграмме рассеяния с треугольником Фанта путем выделения области поиска;
— указанием частотных диапазонов поиска формант;
— по установленным горизонтальным маркам с заданием допуска в герцах и процентах;
— путем поиска всех звуков.

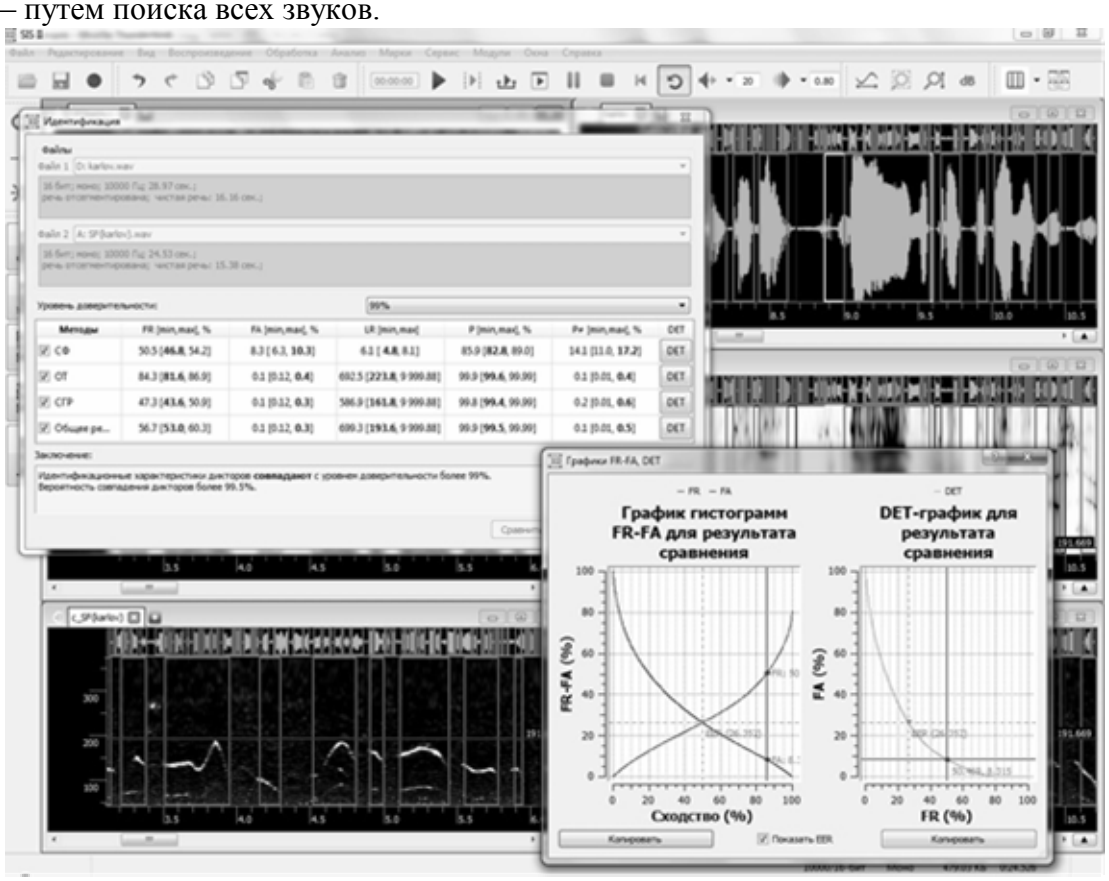


Рис. 1

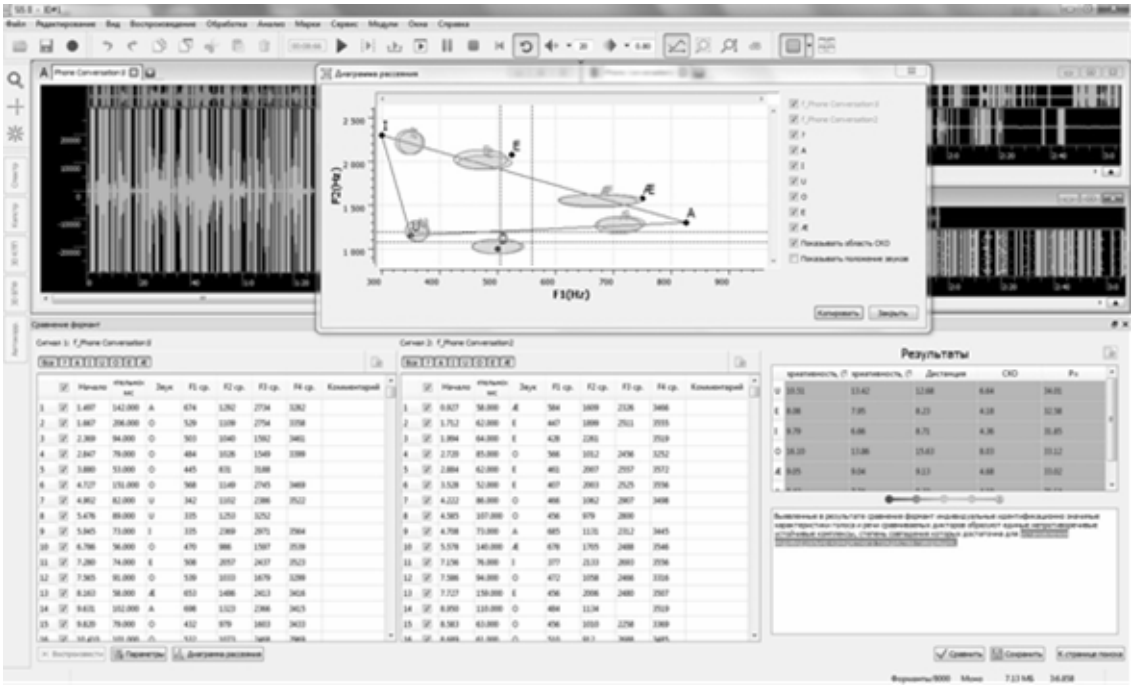


Рис. 2

После того как опорные фрагменты для известного и неизвестного дикторов набраны, эксперт может начать сравнение. Модуль автоматически рассчитывает внутридикторскую и

междикторскую вариативность формантных траекторий для выбранных звуков и принимает решение о положительной/отрицательной идентификации или неопределенном результате.

3. Модуль сравнения основного тона позволяет автоматизировать процесс идентификации дикторов с помощью метода анализа мелодического контура. Метод предназначен для сравнения речевых образцов на основе параметров реализации однотипных элементов структуры мелодического контура. Возможен анализ 18 типов фрагментов контура и 15 параметров их описания, включая значения минимума, максимума, скорости изменения тона, эксцесса, скоса и др. Модуль возвращает результаты сравнения в виде процентного совпадения для каждого из параметров и принимает решение о положительной или отрицательной идентификации или неопределенном результате. Все данные могут экспортироваться в текстовый отчет.

4. Модуль лингвистического анализа позволяет выполнять лингвистическую часть идентификационного исследования и проводить сравнение основного тона на материале русской речи. Он реализован в виде реестра признаков (290 позиций) со звуковыми примерами. Эксперт может отметить три степени выраженности и частотности каждого признака, а также отметить пример его реализации признака на фонограмме. После заполнения таблицы признаков для двух дикторов в модуле лингвистического анализа будет представлена общая статистика по несовпавшим, полностью совпавшим и частично совпавшим признакам.

5. Модуль аудитивного анализа позволяет проводить идентификацию по аудитивным признакам голоса и речи дикторов. Модуль реализован в виде списка из 126 признаков с примерами их звучания. Для каждого признака эксперт может указать три степени его выраженности и частотности, а также отметить пример реализации признака на фонограмме. После заполнения таблицы признаков для двух дикторов в модуле аудитивного анализа можно увидеть общую статистику по несовпавшим, полностью совпавшим и частично совпавшим признакам.

6. Модуль обобщенного решения. Основываясь на результатах работы каждого из модулей, программа позволяет автоматически принять единое идентификационное решение [12]. При этом пользователь имеет возможность изменять границы применимости каждого из отдельных методов исследования, изменяя таким образом его весовой коэффициент в общем решении.

Важными преимуществами специализированного звукового редактора SIS II являются автоматическая оценка качества фонограмм, высокое качество вычисления спектрограмм и автоматического выделения формант. Благодаря уникальному набору функций, обеспечивающих высокую эффективность работы эксперта и расширяющих его возможности при проведении идентификации диктора, SIS II широко используется в отечественных и зарубежных экспертных лабораториях.

Заключение. Анализ имеющихся решений показал, что только в специализированном звуковом редакторе SIS II имеется модуль обобщенного решения, который позволяет автоматизированно с использованием данных, вводимых экспертом, ответить на ключевой вопрос фонографической экспертизы — тождественны ли голоса на сравниваемых фонограммах.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Хитров М. В.* и др. Фоноскопическое исследование фонограмм речи: Исследование достоверности фонограмм. Кн. I. СПб: Юридический центр-Пресс, 2011.

2. *Добрякова М. В.* Зависимость между вербальной спецификой эксперта и надежностью идентификации иноязычного говорящего: Автореф. дис. ... канд. филол. наук. М., 2003.

3. *Kersta L. G.* Voiceprint Identification // Nature. 1962. Vol. 196, N 4861. P. 1253—1257.

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

5. *Harrison P.* Variability of Formant Measurements. Submitted in partial fulfilment of the degree of MA. University of York, UK, 2004.

6. *Enzinger E., Zhang C., Morrison G. S.* Voice source features for forensic voice comparison — an evaluation of the Glottex software package // Proc. of Odyssey. Singapore, 2012. P. 78—85.

7. *Hollien H.* Forensic voice identification. NY, 2002.

8. Программный комплекс „OTExpert“ [Электронный ресурс]: <http://фоноскопическая.рф/forensic_audio-video/forensic_tools/otexpert/>.

9. *Каганов А. Ш.* Криминалистическая экспертиза звукозаписей. М.: Юрлитинформ, 2005.

10. Идентификация лиц по фонограммам русской речи на автоматизированной системе „Диалект“: Пособие для экспертов / Под ред. *А. В. Фесенко*. М., 1996.

11. *Тимофеев И. Н.* Применение автоматизированной системы „Диалект“ на базе компьютерной речевой лаборатории CLS (США) при решении задач идентификации дикторов: Метод. рекомендации. М.: ЭКЦ МВД России, 2000.

12. *Матвеев Ю. Н.* Оценка доверительного интервала общего решения ансамбля классификаторов // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 74—79.

Сведения об авторах

Елена Владимировна Булгакова — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем;
E-mail: bulgakova@speechpro.com

Екатерина Викторовна Краснова — ООО „ЦРТ“, Санкт-Петербург; научный сотрудник;
E-mail: krasnova@speechpro.com

Рекомендована кафедрой
речевых информационных систем

Поступила в редакцию
22.10.13 г.

УДК 004.93+57.087.1

Д. В. Дырмовский, С. Л. Коваль, М. В. Хитров

КОНЦЕПЦИЯ СИСТЕМЫ НАЦИОНАЛЬНОГО ФОНОУЧЕТА И ГОЛОСОВОГО БИОМЕТРИЧЕСКОГО ПОИСКА

Представлена концепция системы национального фоноучета и голосового биометрического поиска.

Ключевые слова: *фоноучет, голосовой биометрический поиск, инфраструктура системы.*

Введение. Многие государства в настоящее время столкнулись с целым рядом задач, связанных с отражением угроз нелегальной эмиграции и контрабанды, предотвращением террористических атак и применением новых подходов при создании систем национальной и международной безопасности. Особую важность в связи с этим приобретает совершенствование систем учета, поиска, выявления и идентификации лиц, нарушивших закон. Решение подобных задач в глобальном масштабе требует применения передовых технологических решений [1].

К инфраструктуре системы национального фоноучета могут быть предъявлены те же требования, что и к инфраструктуре любой информационной системы национального масштаба: экономическая эффективность, с возможностью масштабирования и модернизации (в случае появления более новых технологий) при минимальных дополнительных инвестициях. Соответствовать таким требованиям может система с модульной структурой, отвечающая следующим критериям [2]:

1) достаточно общая и гибкая физическая инфраструктура для совершенствования по мере изменения технологий;

2) возможность интеграции в информационную систему как можно большего количества существующего в различных учреждениях и компаниях оборудования;

3) сетевой сервер размещается таким образом, чтобы было возможным его эффективное управление и обслуживание.

4) использование таких топологий соединений, которые будут как экономически эффективными, так и управляемыми.

Принципы построения систем. Основными задачами систем национального фоноучета и голосового биометрического поиска являются:

— централизованный сбор и хранение образцов речи и голосовых биометрических признаков личностей, подлежащих учету;

— хранение сопутствующей информации — фотоизображений, установочных данных и любых других данных (метаданных), характеризующей личность или обстоятельства получения образца речи;

— поиск по биометрическим признакам и сопровождающей информации дикторов по всей базе данных или ее части;

— автоматическая идентификация/верификация с целью определения или подтверждения личности стоящего на учете лица.

Система национального фоноучета и голосового биометрического поиска, способная решить множество задач в области правоохранительной деятельности, должна быть построена на следующих принципах:

1) *территориальная распределенность* — возможность вести учет и осуществлять поиск по банкам системы из любого подразделения правоохранительных органов;

2) *открытость и консолидация данных* — возможность добавлять в систему банки и базы речевых данных, содержащие иную информацию, например, цифровые модели других биометрических признаков (лица, отпечатков пальцев, сетчатки глаза, ДНК) или учетные личные карточки подозреваемых;

3) *безопасность данных* — должен быть реализован контроль и разграничение доступа к данным системы на базе голосовой биометрии (возможно использование традиционных систем);

4) *автоматизация рабочих процессов* — в функционирование системы должен быть вовлечен каждый пользователь согласно его компетенциям через гибкие механизмы полей, групп и рабочих процессов.

Территориальная распределенность системы. Сбор биометрической информации у населения также должен производиться непосредственно на местах (муниципальные образования, организации).

Для решения этой задачи необходимо, чтобы:

— был обеспечен доступ к системе всех пользователей по защищенному протоколу HTTPS, сочетающему криптографическую стойкость и повсеместную доступность как в локальных и ведомственных сетях, так и в сети Интернет;

— система не имела технических ограничений ни по объему хранимой информации, ни по числу пользователей, которые получают доступ к ресурсам центрального хранилища и поисковой машины удаленно;

— обеспечивались web-интерфейс и поддержка большинства браузеров в совокупности с защищенными протоколами передачи данных для высокой мобильности специалиста, ведущего расследование. Благодаря этому можно будет образцы биометрических данных передавать в систему непосредственно с места совершения преступлений и там же получать результаты поиска.

На рис. 1 представлена схема территориального распределения пользователей системы.

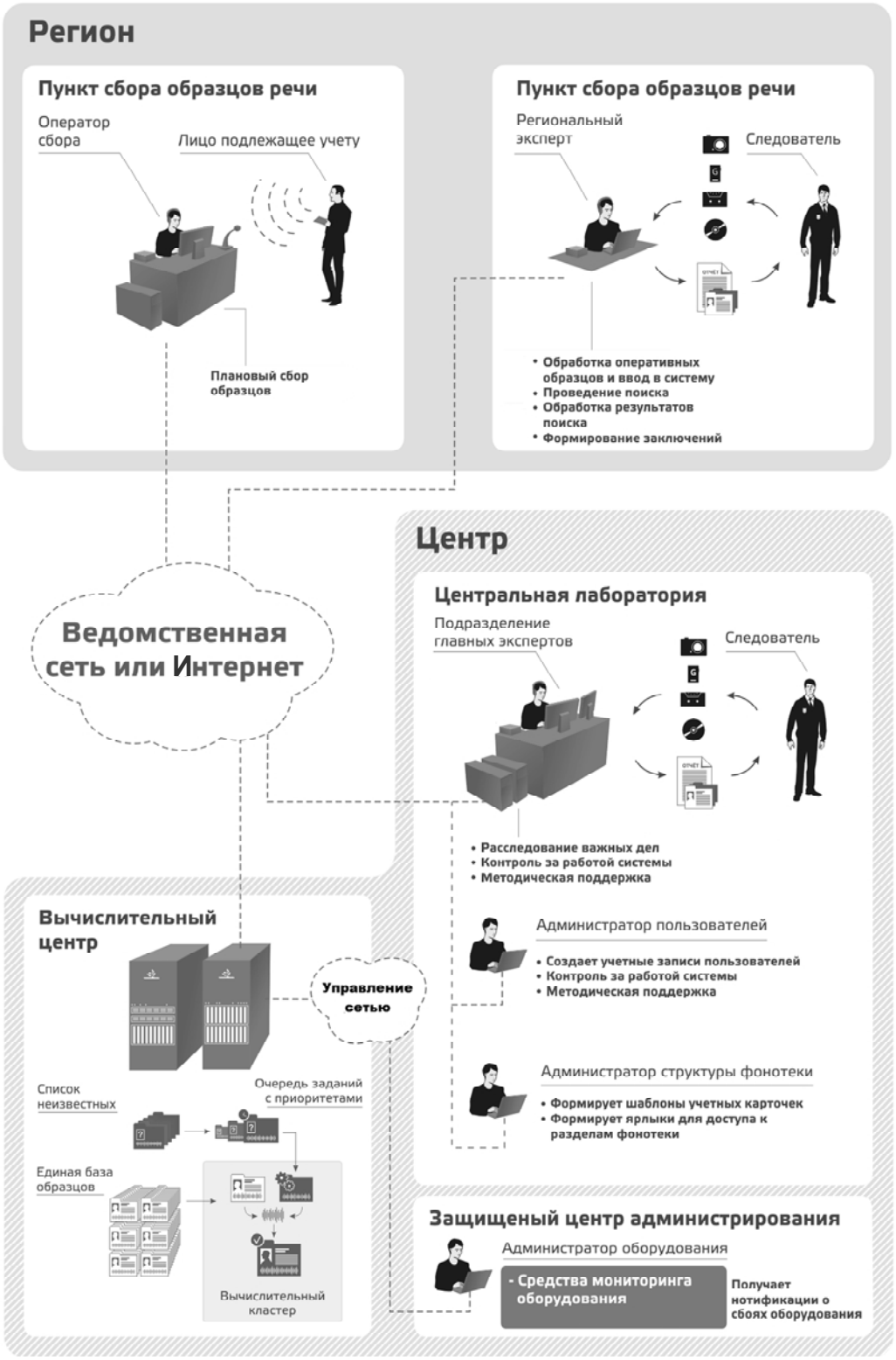


Рис. 1

Функциональные уровни системы. Система голосового биометрического поиска и национального криминалистического фоноучета включает три уровня.

1. *Сбор данных, подлежащих криминалистическому учету*: учетные карточки разыскиваемых лиц и образцы их речи, пригодные для идентификации.

2. *Хранение и обработка данных, осуществление вычислительных операций*: хранение банка биометрических данных, база данных учетных карточек, осуществление непосредственно идентификации и поиска по банку данных с помощью мощного вычислительного кластера.

3. *Операции пользователя*: постановка заданий на поиск, определение параметров поиска, вывод результатов в форме отчетов, оперативная нотификация сотрудников в случае обнаружения разыскиваемого лица в каналах связи или банках фонограмм, пополняемых в процессе оперативно-розыскных мероприятий, задания экспертам на обработку и анализ результатов поиска и фонограмм.

Уровень сбора данных базируется на существующих в правоохранительных органах технологиях сбора и учета криминалистических данных: результаты проведения оперативно-розыскных мероприятий (в том числе с помощью СОРМ, скрытых средств записи переговоров и т.д.); аудиозаписи приема вызовов дежурными частями территориальных подразделений; аудиозаписи дачи показаний; судебных разбирательств; данные экспертного фоноскопического учета.

Кроме того, в соответствии с законодательством возможна организация планового сбора образцов речи у населения: заключенных в тюрьмах; задержанных, находящихся в полицейских участках; сотрудников правоохранительных органов; населения в точках оформления удостоверений личности.

Уровень обработки данных должен опираться на возможности современных технологий голосовой биометрической идентификации личностей и мировой опыт внедрения систем фоноучета различного масштаба — от экспертной лаборатории до национальной системы. Выделим основные характеристики, которыми должна обладать система голосового биометрического поиска и национального криминалистического фоноучета:

— высокая надежность поиска/идентификации по голосу (не менее 97 % правильной идентификации) с использованием языко- и текстонезависимых биометрических методов;

— высокая скорость поиска (поиск из 20 000 образцов за 3—10 мин при одновременной обработке не менее 100 запросов);

— надежная защита передачи и хранения информации (использование протокола HTTPS, распределение решения по слоям безопасности на аппаратном уровне);

— быстрота развертывания (возможность доступа по любым сетям с использованием защищенных протоколов, наличие web-интерфейса, автоматическое обновление клиентского программного обеспечения);

— работа в режиме высокой готовности 24×7 (автоматическое перераспределение нагрузки в случае выхода оборудования из строя, резервирование на всех уровнях, возможность обновления программного обеспечения без остановки работы, дистанционный мониторинг оборудования);

— эффективная работа с большим числом пользователей (автоматическая балансировка нагрузки, кластерные параллельные вычисления, очередь заданий с приоритетами);

— масштабируемость с возможностью расширения системы без остановки работы (от рабочей станции для нужд подразделения до вычислительного центра для национальной системы).

Уровень операций пользователя должен обеспечивать гибкую схему рабочих процессов и ролей пользователей. Каждый пользователь согласно своей роли и положению в иерархии групп, с одной стороны, получает доступ к необходимой для его работы информации, а с

другой — привлекается к решению тех задач, в которых он наиболее компетентен. Таким образом, высокая степень автоматизации и простота использования позволяют обучать пользователей непосредственно в процессе эксплуатации системы, а общие для всех пользователей средства (оповещения по e-mail, настраиваемые рабочие процессы) позволяют системе эффективно решать правоохранительные задачи.

Рабочие процессы системы. Работа любой информационной системы национального масштаба невозможна без разделения ролей пользователей, автоматизации рабочих процессов и аудита, что обеспечивает такие свойства системы, как безопасность выполняемых пользователями операций и эффективность работы каждого пользователя вне зависимости от его квалификации.

Основными сценариями использования системы криминалистического учета являются помещение образцов в базу и поиск образцов по степени сходства.

Можно выделить различные логические типы поиска:

1) поиск „известного среди известных“:

— проверка постановки на учет уже имеющегося в базе диктора (исключение постановки на учет диктора под чужим именем);

2) поиск „неизвестного среди известных“ и наоборот:

— установление личности интересующего диктора;

3) поиск „неизвестного среди неизвестных“:

— проверка принадлежности образцов дикторов одному или нескольким людям;

— проверка причастности диктора к нескольким преступлениям.

Рабочие процессы системы должны обеспечивать максимальную эффективность и оперативность всех перечисленных типов поиска.

Существенной сложностью при внедрении систем фоноучета является подготовка персонала. При внедрении систем на уровне лабораторий все пользователи должны иметь квалификацию экспертов-фоноскопистов, подготовка которых является долгим и дорогим процессом. Внедрение системы национального масштаба с высоким уровнем автоматизации рабочих процессов позволяет преодолеть эту проблему за счет повышения квалификации оператора системы, от выполнения самых простых задач до уровня эксперта. Оператор обучается постепенно, выполняя практические задачи уже с первых дней работы.

Рекомендуется выделять следующие роли пользователей системы:

— оператор сбора образцов — производит плановый сбор образцов речи и отвечает только за качественную запись и соответствие учетных данных известной личности;

— оператор проверки образцов — анализирует качество записанных или полученных в результате оперативно-розыскной деятельности образцов и при необходимости сегментирует их;

— оператор поиска — выполняет несложные поисковые запросы, требующие немедленного ответа.

— эксперт — анализирует результаты поиска для предоставления их в суд.

Квалифицированные пользователи могут совмещать несколько ролей.

Схемы рабочих процессов разрабатываются в зависимости от возможностей и потребностей ведомства, внедряющего эту систему. Рабочие процессы не являются фиксированными во времени и могут эволюционировать вместе с системой. При внедрении может быть выбрана наиболее простая схема, а по мере вовлечения в работу с системой новых пользователей могут вводиться как новые роли, так и новые схемы рабочих процессов.

На рис. 2, 3 приведены схемы жизненного цикла учетной карточки в системе криминалистического учета, отражающие автоматизацию рабочих процессов. В данных примерах учетные карточки известных и неизвестных личностей обрабатываются по разным схемам рабочих процессов.

Система обеспечивает высочайшую степень автоматизации. Таким образом, любая учетная карточка, введенная в систему, будет обработана (с привлечением экспертов, если это требуется) и использована для получения новой информации, полезной при расследовании или в оперативно-розыскных мероприятиях.

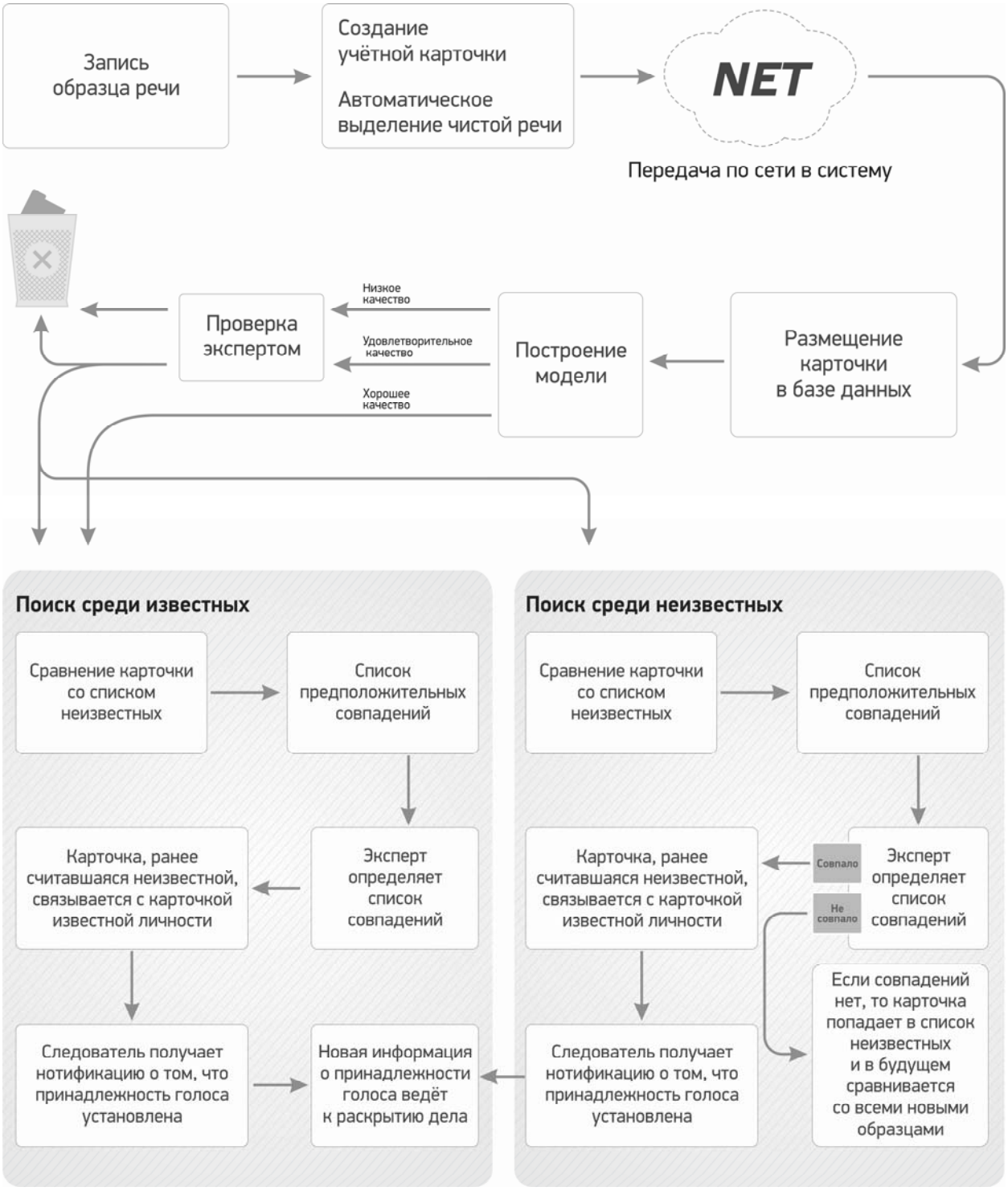


Рис. 2

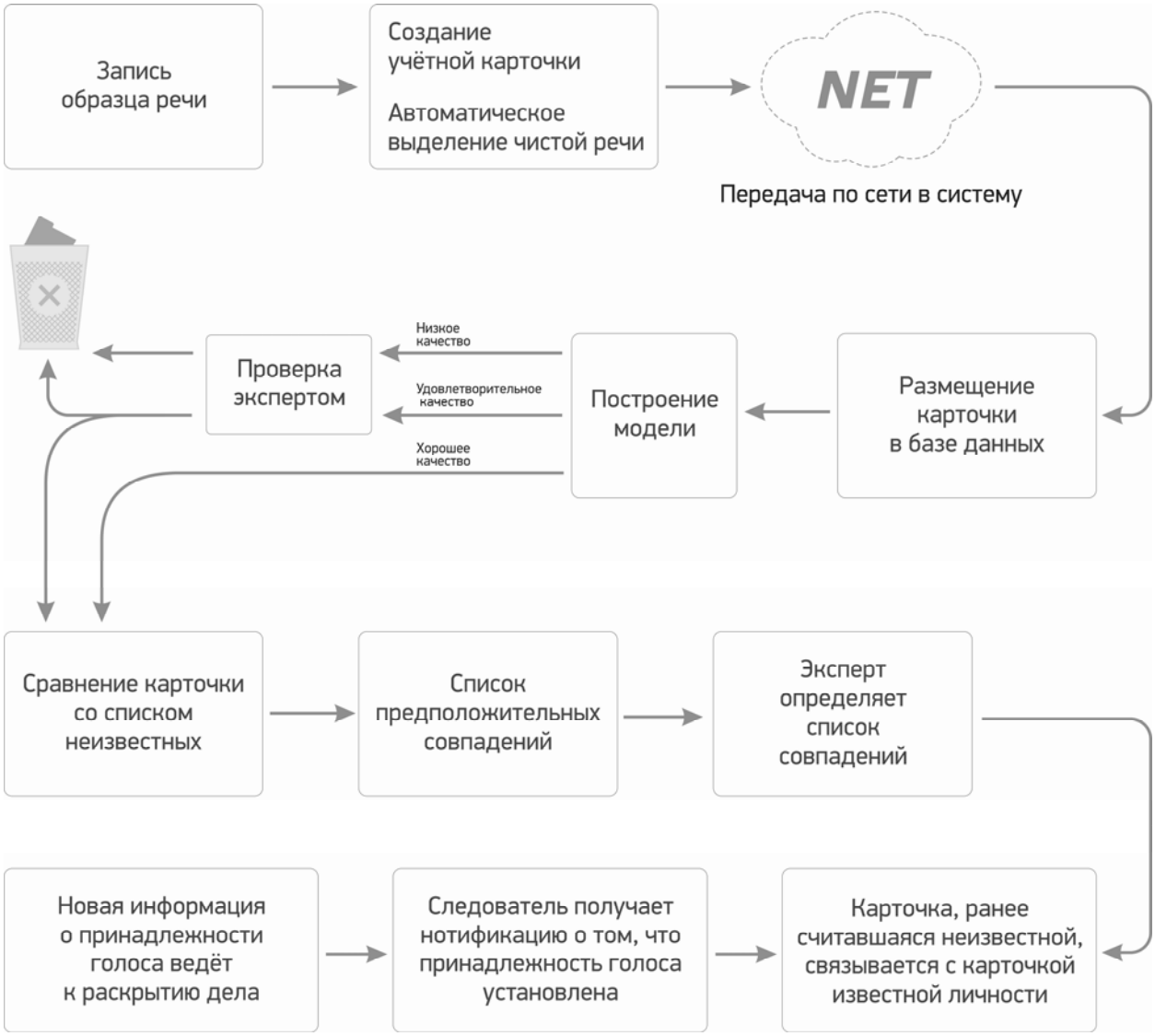


Рис. 3

Автоматизация рабочих процессов обеспечивает максимально оперативную обработку речевого материала, являющегося уликой.

Заключение. В работе предложена концепция построения системы национального фоноучета и голосового биометрического поиска, положенная в основу реальных систем, развернутых ООО „ЦРТ“ в различных странах (Мексика, Эквадор) и ведомствах (МВД, ФСКН и др.).

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

СПИСОК ЛИТЕРАТУРЫ

1. *Настасенко М. В., Дырмовский Д. В.* Эффективное использование речевой информации и биометрических технологий в силовых структурах — научно-технические решения // Сб. тр. XX Междунар. науч. конф. „Информатизация и информационная безопасность правоохранительных органов“. М., 2011. С. 356—359.

2. *Nair R. R.* Strategic plans for designing information systems under local government // Proc. 18th National IASLIC Seminar. Thrissur, Kerala: Agricultural University, 1998. P. 11.

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

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

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

Сведения об авторах

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

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 004.93+57.087.1

Ю. Н. Матвеев, А. К. Шулипа

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Таблица 1

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

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

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

Таблица 2

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

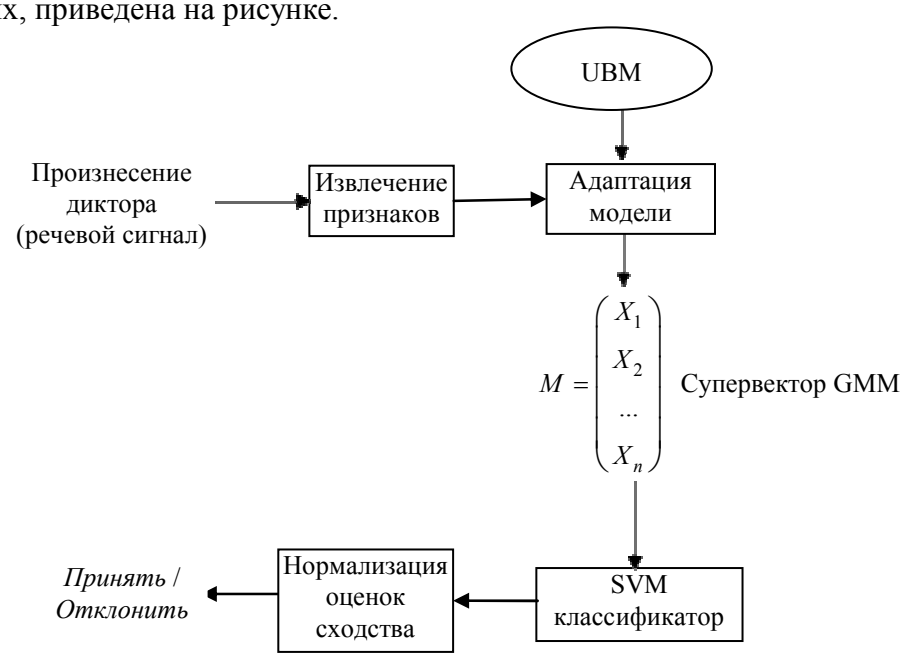
Число дикторов	Метод			
	диффузных карт		GMM	*k*-NN
	несимметричный	симметричный		
2	98,2	99,1	97,4	97,9
3	97,8	98,7	95,2	97,8
5	96,5	96,0	92,1	93,4
10	92,5	91,9	87,4	89,0
20	86,4	84,6	83,3	84,5

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

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

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

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



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

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

Для экспериментальных исследований использовалось несколько речевых баз:

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

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

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

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

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

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

2) Супервекторы конфигурировались в виде матриц:

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

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

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

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

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

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

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

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

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

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

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

— базовая;

— схема с нелинейной редукцией пространства GMM супервекторов методом Isomap.

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

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

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

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

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

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

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

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

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

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

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

СПИСОК ЛИТЕРАТУРЫ

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

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

12. *Reynolds D.A., Quatieri T.F., Dunn R.B.* Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing. 2000. Vol. 10. P. 19—41.

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

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

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

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

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

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

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

Сведения об авторах

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 57.087.1

С. А. Новоселов, В. А. Сухмель, А. В. Шолохов, Т. С. Пеховский

ПРИМЕНЕНИЕ DTW-МЕТОДА ДЛЯ МУЛЬТИСЕССИОННОГО ОБУЧЕНИЯ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ В ЗАДАЧЕ ТЕКСТОЗАВИСИМОЙ ВЕРИФИКАЦИИ ДИКТОРА

Представлен метод обучения скрытых марковских моделей по нескольким вариантам произнесения парольной фразы с помощью алгоритма динамического временного выравнивания сигналов. Метод позволяет создавать точные статистические модели речевых сигналов и снижать вероятность возникновения ошибок верификации.

Ключевые слова: *текстозависимая верификация диктора, MFCC, HMM, GMM, DTW.*

Введение. В последнее время в области распознавания диктора по голосу преобладают исследования текстонезависимых методов [1—8]. Основу всех алгоритмов идентификации личности по голосу составляет GMM-UBM-подход, основанный на совместном использовании смесей гауссовых распределений (Gaussian Mixture Model, GMM) и универсальной фоновой модели (Universal Background Model, UBM) [9]. Определенного успеха в этой области удалось достичь за счет использования методов снижения размерностей и различных классификаторов для параметров средних (супервекторов) GMM-моделей дикторов [1—3, 6, 8]. Однако эффективность таких систем зависит от длительности речевых сигналов [4, 5, 7].

Одним из вариантов реализации систем верификации диктора по голосовому паролю может быть комбинирование дикторонезависимого распознавателя речи и текстонезависимого распознавателя диктора. Такой подход является языкозависимым и требует достаточного количества ресурсов. На наш взгляд, перспективен подход, при котором на этапе обучения системы верификации создается статистическая модель речевого сигнала, способная описывать последовательность звуков парольной фразы, а также особенности их произнесения конкретным диктором. Возможность создания такой скрытой марковской модели (Hidden Markov Model, HMM) подтверждается результатами работы [10]. В ней показано, что эмиссионные распределения состояний модели HMM-GMM парольной фразы могут быть аппроксимированы на основе адаптации статистической модели голоса диктора, которая, в свою очередь, обучается классическими методами текстонезависимой идентификации [1—9]. Верификационный тест при использовании такой генеративной модели удобно проводить путем оценки логарифма правдоподобия для тестового речевого сигнала с учетом наиболее вероятной последовательности состояний, т.е. с учетом пути Витерби [11].

В настоящей статье исследуется иерархическая система текстозависимой верификации диктора на различных базах речевых данных и разрабатывается метод обучения HMM-GMM-модели парольной фразы при наличии нескольких вариантов произнесения (сессией), на основе метода динамического программирования DTW (Dynamic Time Warping) [12].

Иерархическая система верификации диктора. В основе иерархического подхода к обучению статистических моделей лежит трехуровневая структура представления акустических признаков речевых сигналов. Подход предполагает последовательную адаптацию моделей по критерию максимума апостериорной вероятности (MAP-адаптация). Для описания парольной фразы используется скрытая марковская модель, что обусловлено возможностью „запоминания“ временной структуры речи с помощью введения множества состояний модели и матрицы вероятностей переходов между этими состояниями [10—14]. Чтобы учесть лингвистическую информацию, которая содержится в голосовом пароле диктора, статистические

модели состояний HMM парольной фразы предложено строить путем адаптации заранее обученной текстонезависимой GMM-модели диктора. Эмиссионные плотности распределения вероятностей марковской модели в данном случае аппроксимируются смесями гауссовых распределений. При таком обучении HMM сохраняются достоинства схемы GMM-UBM, которая успешно применяется в решении задач текстонезависимого распознавания диктора.

Общее описание. На рис. 1 схематически представлен иерархический подход к обучению HMM. Все узлы этой схемы являются смесями гауссовых распределений. Верхний уровень системы — это универсальная фоновая модель, которая статистически описывает общее акустическое пространство признаков различных голосов [9].

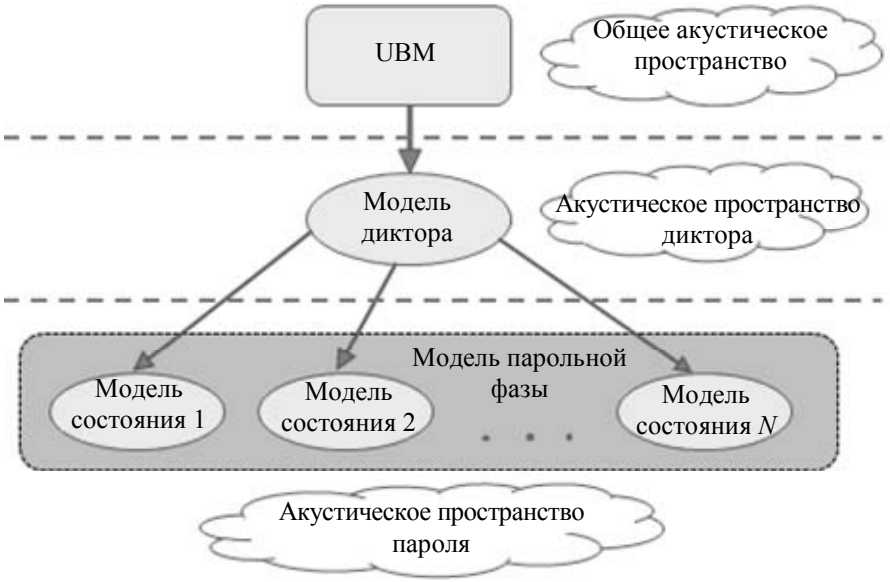


Рис. 1

Средний уровень системы предназначен для моделирования текстонезависимой информации о голосе конкретного человека. Параметры распределения акустических характеристик голоса могут быть оценены на основе критерия максимума апостериорной вероятности по известной универсальной фоновой модели. На этом уровне адаптируются только векторы математических ожиданий гауссиан, а все остальные параметры заимствуются у модели UBM. Модель голоса диктора используется далее на текстозависимом уровне для построения HMM-модели. Статистические модели состояний MM обучаются путем адаптации дикторской модели GMM, полученной на среднем уровне, к универсальной фоновой. При этом трансформируются только весовые коэффициенты гауссиан. Так формируется HMM-GMM-модель, которая описывает особенности голоса диктора и временную структуру последовательности звуков парольной фразы. Рассмотрим подробнее этапы обучения системы.

Обучение системы. Алгоритм обучения HMM-GMM состоит из трех этапов. На первом строится универсальная фоновая модель верхнего уровня иерархической системы. Расчет модели проходит на большом объеме речевой информации с помощью классического алгоритма максимизации ожидания (expectation maximization, EM) [1, 9].

Обучение дикторской модели текстонезависимого уровня осуществляется путем адаптации UBM-модели с использованием речевых сигналов конкретного диктора. На этом этапе необходимо учитывать решение детектора речевой активности для определения речевых сегментов сигналов, поскольку именно они необходимы для построения адекватной статистической модели голоса.

На последнем этапе обучается итоговая модель HMM (рис. 2). По мнению авторов работ [10, 12], на этом уровне необязательно исключать неречевые сегменты из парольной фразы, поскольку HMM способна моделировать паузы в речи. Весь речевой сигнал разбивается на

участки одинаковой длины. Каждое состояние HMM обучается методом адаптации дикторской модели верхнего уровня по данным соответствующего сегмента. MAP-адаптация выполняется только для весовых коэффициентов гауссиан дикторской модели. После того как модели состояний построены, HMM с равновероятной матрицей переходов оптимизируется с помощью классического алгоритма Витерби, а вероятности переходов между состояниями пересчитываются пропорционально относительной длине смежных сегментов.

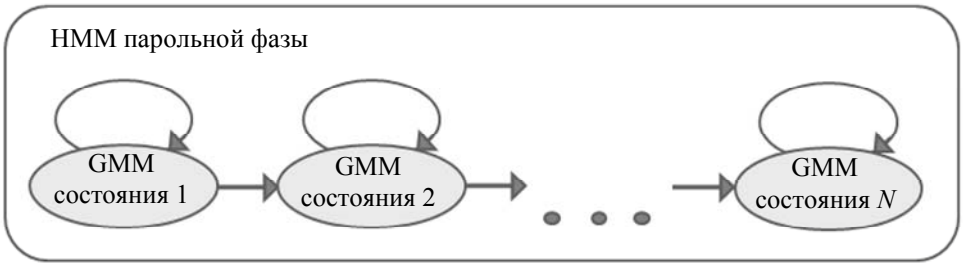


Рис. 2

Тестирование. Для определения метрик сравнения при проведении верификационных тестов сначала рассчитывается путь Витерби для тестовой последовательности речевых признаков. В декодировании Витерби участвуют все сегменты сигнала без исключения. Логарифм правдоподобия HMM λ вычисляется с учетом пути Витерби по состояниям и может быть записан как сумма логарифмов правдоподобия моделей состояний λ_S парольной фразы на речевых и неречевых кадрах:

$$\log p\left(X|\lambda\right) = \sum_{S\in\text{Speech}} \log p\left(X_S|\lambda_S\right) + \sum_{S\notin\text{Speech}} \log p(X_S \mid \lambda_S),$$

где X — параметры речевого сигнала, X_S — параметры, соответствующие состоянию S, Speech — множество состояний парольной фазы, содержащих речь.

Финальная метрика сравнения score формируется только из логарифмов правдоподобия на речевых сегментах путем нормализации с помощью логарифма правдоподобия универсальной фоновой модели λ_{UBM} и усреднения по всем таким состояниям S:

$$\text{score} = \frac{1}{S} \sum_{S\in\text{Speech}} \left[\log p\left(X_S|\lambda_S\right) - \log p(X_S \mid \lambda_{\text{UBM}})\right]$$

При наличии нескольких вариантов произнесения парольной фразы неясно, каким образом проводить обучение HMM-GMM-модели диктора, используя всю доступную речевую информацию. Для среднего уровня иерархической системы этот вопрос легко разрешается путем простой конкатенации информативных параметров нескольких произнесений парольной фразы. Таким образом, возможно получить более точно соответствующую дикторскую модель на текстонезависимом уровне [10]. Использование техники динамического временного выравнивания сигналов [15, 16] для обучения последнего уровня иерархии позволит принимать во внимание вариативность произнесения отдельных звуков парольной фразы и более точно формировать статистические модели состояний HMM с помощью критерия MAP по нескольким произнесениям.

Динамическое временное выравнивание сигналов. Целью DTW [16] является сравнение двух последовательностей $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ и $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, где x_n и y_n — векторы временных параметров сигналов (например, мел-частотные кепстральные коэффициенты, MFCC [17]). Для того чтобы сравнить два вектора $\mathbf{x}$ и $\mathbf{y}$, введем понятие локальной дистанции:

$$d\left(\mathbf{x},\mathbf{y}\right) = \text{Norm}\left(\mathbf{x},\mathbf{y}\right) \to \left[0,+\infty\right).$$

С использованием локальной метрики сравнения векторов вычислим матрицу:

$$\mathbf{C}_{n,m} = d\left(\mathbf{x}_n, \mathbf{y}_m\right),$$

по которой определяется оптимальный путь сравнений последовательностей векторов **x** и **y**. Определив путь, можно проводить временную темпокоррекцию двух сигналов и определять участки речи, наиболее соответствующие друг другу по критерию выбранной метрики.

Обучение моделей состояний. Рассмотрим случай обучения GMM-моделей состояний при наличии двух вариантов произнесения парольной фразы с помощью DTW. Вначале речевые сигналы выравниваются по времени. Временная шкала первого сигнала, так же как и в базовом подходе [10], разбивается на равные сегменты. Данные для обучения статистических моделей состояний теперь будут формироваться по двум последовательностям векторов информативных параметров разных сигналов (рис. 3). Каждое состояние HMM обучается методом адаптации модели диктора по вновь сформированным данным соответствующего сегмента пароля.

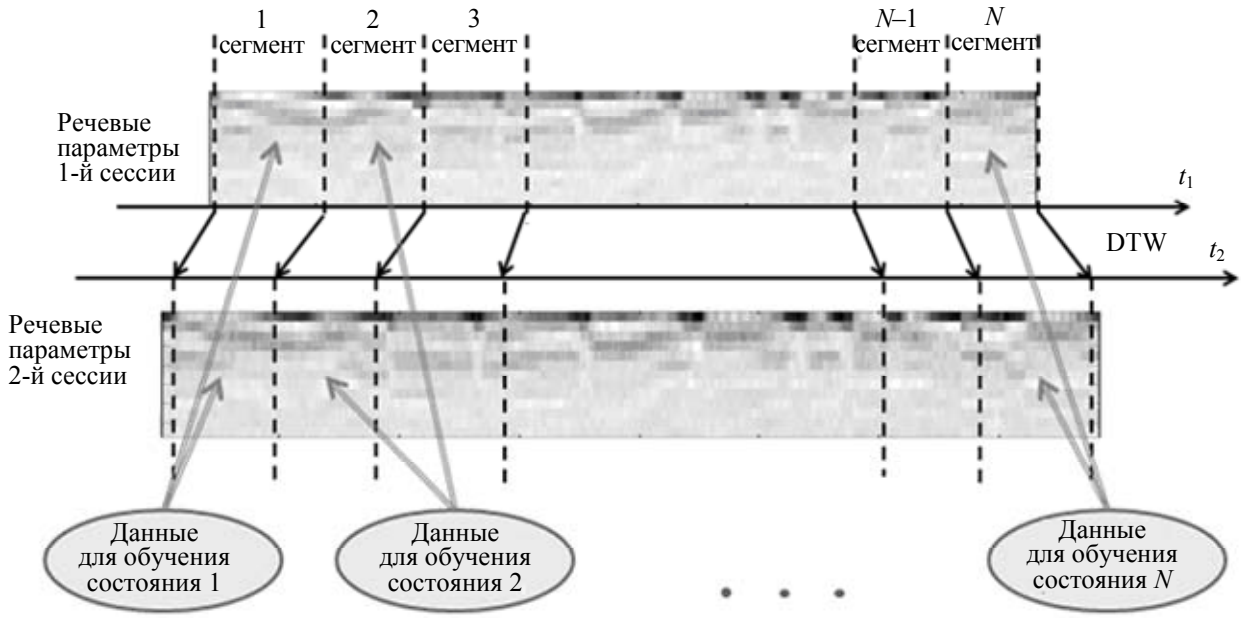


Рис 3

На этапе оптимизации моделей расчет пути Витерби производится только для первого речевого сигнала. В результате получается новое разбиение временной шкалы и происходит переобучение статистических моделей состояний. Формирование матрицы вероятностей переходов между состояниями осуществляется по правилам, описанным выше.

С помощью метода DTW аналогичным образом можно проводить обучение и для большего количества вариантов произнесения парольной фразы.

Параметры системы тестирования

Информативные параметры речевых сигналов, для их получения используются MFCC. Размерность векторов равна 13, предполагаются, что компоненты векторов некоррелированны между собой.

Детектор речевой активности построен на основе GMM-моделей.

Универсальная фоновая модель. Исходя из результатов статьи [10] размерность UBM выбрана равной 64. Для баз данных *ФИО_цифры_0_6* и *Цифры_0_9* обучалась на множестве 27 дикторов, для базы *POLYCOST* — строилась по рекомендациям работы [18].

Длина сегментов для обучения состояний HMM. Исследовалась зависимость равновероятной ошибки первого и второго рода (Equal Error Rate, EER) системы верификации дикторов

от длины сегментов, на которые разбивается парольная фраза в процессе обучения HMM (рис. 4). Модели обучались по базовому правилу на одном произнесении пароля. Тестирование проводилось по всем выбранным базам. Минимальное значение EER для трех случаев (*1 — Цифры_0_9*, *2 — ФИО_цифры_0_6*, *3 — POLYCOST*) достигается при сегментах порядка 0,03—0,04 с. Это означает, что на начальном этапе обучения HMM целесообразно разбивать речевые сигналы на одинаковые участки длительностью около 35 мс.

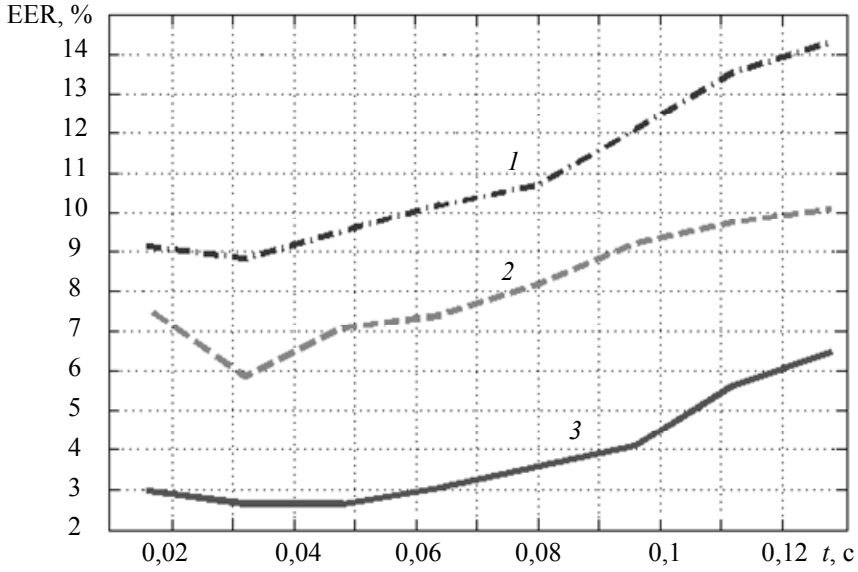


Рис. 4

Экспериментальные исследования выполнены на русскоязычных речевых базах *ФИО_цифры_0_6*, *Цифры_0_9* и англоязычной части базы *POLYCOST*. В настоящей работе эксперименты проводились на голосах мужских дикторов.

— Текстозависимая база *ФИО_цифры_0_6* содержит аудиозаписи 127 мужских голосов, которые произносят свои фамилию, имя, отчество и последовательность цифр от 0 до 6. База содержит 15 сессий каждого диктора. Интервал между записями сессий — не менее 1 дня и не более одной недели. Записи проводились в GSM-канале при использовании разных сотовых телефонов.

— База *Цифры_0_9* включает в себя записи 127 мужских дикторов (тех же, что в *ФИО_цифры_0_6*), которые произносят общую фразу, состоящую из последовательности цифр от 0 до 9. Условия сбора речевой базы те же, что и для *ФИО_цифры_0_6*.

— База *POLYCOST* [19] содержит 10 сессий, записанных 134 дикторами из 14 стран. Каждая сессия состоит из 14 произнесений; 4 повторений кода из 7 цифр, 5 последовательностей из 10 цифр, 2 фиксированных фраз, 1 международного номера телефона и 2 предложений с речью на родном языке диктора. Как и рекомендовано в [18], были исправлены ошибки базы и для проведения исследований тектозависимой верификации диктора выбрана парольная фраза “Joe took father's green shoe bench out”.

Результаты исследований EER рассматриваемой системы верификации приводятся для следующих вариантов обучения модели:

I — обучение дикторской модели и модели состояний по одному произнесению пароля;

II — обучение дикторской модели по двум произнесениям, моделей состояний — по одному из произнесений;

III — обучение дикторской модели и моделей состояний по двум произнесениям, с применением выравнивания MFCC-последовательностей по времени.

Из табл. 1 видно, что при обучении HMM-GMM-моделей по нескольким вариантам произнесения парольной фразы применение DTW-метода позволяет снизить уровень равновероятной ошибки верификации. Видно также, что при обучении по двум произнесениям только

дикторской модели второго уровня EER системы верификации в некоторых случаях увеличивается по сравнению с EER системы, обученной на одном произнесении. Это может быть следствием получения менее адекватных моделей состояний HMM при таком варианте обучения.

Таблица 1

Сравнение EER (%) для разных вариантов обучения

База	I	II	III		
			L_1	L_2	cos
ФИО_цифры_0_6	5,85	6,61	5,16	4,09	4,25
Цифры_0_9	8,81	9,08	7,40	7,30	7,53
POLYCOST	2,63	2,58	2,63	1,76	1,69

Дополнительно исследовалось влияние выбора локальной метрики сравнения DTW-методе на ошибки системы верификации. Выявлено, что норма схожести MFCC-векторов является лучшей по сравнению с нормой L_2 и косинусной метрикой [20].

При сравнении значений EER, полученных на базах *ФИО_цифры_0_6* и *Цифры_0_9*, необходимо учесть, что при работе с первой моделируется случай, когда речевой пароль не известен злоумышленнику, а вторая — когда злоумышленник знает парольную фразу. Голоса дикторов и условия записи аудиофайлов в этих базах совпадают. Из результатов табл. 1 видно, что значение EER в случае неизвестного пароля в среднем на 3 % меньше, чем при известном пароле.

Результаты, полученные на базе *POLYCOST*, лучше, поскольку она меньше по объему и речевые сигналы в ней менее искажены каналом передачи.

В табл. 2 приведены значения EER, полученные при тестировании системы верификации обученной на пяти вариантах произнесения парольной фразы. В методе DTW для сравнения векторов признаков использовалась норма L_2. Ошибки верификации снижаются при увеличении объема обучающих данных. Необходимо отметить, что для текстозависимой базы *ФИО_цифры_0_6* при обучении моделей на четырех вариантах произнесения пароля наблюдается снижение EER на 40 % по сравнению со случаем обучения на одном варианте произнесения. При увеличении числа произнесений до 5 относительное изменение EER составляет ≈47 %. Для баз *Цифры_0_9* и *POLYCOST* при использовании для обучения 5 сессий вместо одной удается снизить ошибку EER на 23 и 50 % соответственно.

Таблица 2

Сравнение EER (%) при обучении на пяти вариантах произнесения

База	1	2	3	4	5
ФИО_цифры_0_6	5,85	4,09	3,92	3,50	3,04
Цифры_0_9	8,81	7,30	7,29	6,92	6,77
POLYCOST	2,63	1,76	1,70	1,32	1,31

Выводы. В работе исследована иерархическая система текстозависимой верификации диктора на трех различных речевых базах. Найдена оптимальная по критерию равновероятной ошибки системы верификации длина сегментов, на которые необходимо разбивать речевой сигнал при обучении HMM-GMM-моделей. Представлен новый метод обучения HMM-GMM-модели парольной фразы при наличии нескольких вариантов произнесения с помощью временной темпокоррекции речевых сигналов. Показано, что в качестве локальной метрики сравнения MFCC-векторов в методе DTW эффективно использовать норму L_2. Анализ результатов показал, что для тектозависимого случая при использовании четырех вариантов произнесения пароля на этапе обучения HMM-GMM вместо одного удается снизить EER системы верификации на 30 %.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Kenny P., Boulianne G., Ouellet P., Dumouchel P.* Speaker and Session Variability in GMM-Based Speaker Verification // IEEE Transact. on Audio, Speech, and Language Processing. 2007. Vol. 15, N 4. P. 1448—1460.

2. *Vogt R. J., Lustri C. J., Sridharan S.* Factor Analysis Modelling for Speaker Verification with Short Utterances // Proc. Of Speaker and Language Recognition Workshop "Odyssey-2008". Stellenbosch, South Africa, 2008. P. 1—5.

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

4. *Vogt R., Sridharan S., Mason M.* Making confident speaker verification decisions with minimal speech // IEEE Transact. On Audio Speech, and Language Processing. 2010. Vol. 18, N 6. P. 1182—1192.

5. *McLaren M., Vogt R., Baker B., Sridharan S.* Experiments in SVM-based Speaker Verification Using Short Utterances // Proc. of Speaker and Language Recognition Workshop (Odyssey—2010). Brno, Czech Republic, 2010. P. 83—90.

6. *Kanagasundaram A., Vogt R., Dean D. B., Sridharan S., Mason M. W.* I-vector based speaker recognition on short utterances // Proc. of 12th Annual Conf. of International Speech Communication Association (INTERSPEECH 2011). Firenze Fiera, Florence, 2011. P. 2341—2344.

7. *Kanagasundaram A., Vogt R. J., Dean D. B., Sridharan S.* PLDA based speaker recognition on short utterances // Proc. of Speaker and Language Recognition Workshop "Odyssey-2012". Singapore, 2012. P. 28—33.

8. *Stafylakis T., Kenny P., Senoussaoui M., Dumouchel P.* PLDA using Gaussian Restricted Boltzmann Machines with application to Speaker Verification // Proc. Of 13th Annual Conf. of Intern. Speech Communication Association (INTERSPEECH 2012). Portland, Oregon, USA. 2012. P. 2341—2344.

9. *Reynolds D., Quatieri T., Dunn R.* Speaker Verification using Adapted Gaussian Mixture Models // Digital Signal Processing. 2000. Vol. 10. P. 19—41,

10. *Larcher A. O., Bonastre J.-F., Mason J. S. D.* From GMM to HMM for embedded password-based speaker recognition // Proc. 16th Europ. Signal Processing Conf. (EUSIPCO-2008). Lausanne, Switzerland, 2008. P. 1—5.

11. *Juang B. H., Rabiner L. R.* Hidden Markov Models for Speech Recognition // Technometrics. 1991. Vol. 33, N 3. P. 251—272.

12. *Geppener V. V., Simonchik K. K., Haidar A. S.* Design of speaker verification systems with the use of an algorithm of Dynamic Time Warping (DTW) // Pattern Recognition and Image Analysis. 2007. Vol. 17, N 4. P. 470—479.

13. *Larcher A. O., Bonastre J.-F., Mason J. S. D.* Reinforced Temporal Structure Information for Embedded Utterance-Based Speaker Recognition // Proc. of 9th Annual Conf. of Intern. Speech Communication Association (INTERSPEECH 2008). Brisbane, Australia, 2008. P. 371—374.

14. *Subramanya A., Zhengyou Z., Surendran A. C., Nguyen P., Narasimhan M., Acero A.* A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification // Proc. of the Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP—2007). Honolulu, Hawaii, USA, 2007. Vol. 4. P. 225—228.

15. *Винцюк Т. К.* Распознавание слов устной речи методами динамического программирования // Кибернетика. 1968. № 1. С. 81—88.

16. *Muller M.* Information Retrieval for Music and Motion. Springer, 2007. 318 p.

17. *Матвеев Ю. Н.* Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56, № 2. С. 47—51.

18. *Melin H., Lindberg J.* Guidelines for experiments on the POLYCOST База. KTH/Centre for Speech Technology [Электронный ресурс]: <http://www.speech.kth.se/cost250/polycost/be/v2.0/>.

19. *Petrovska D., Hennebert J., Melin H., Genoud D.* Polycost: A Telephone-Speech Database for Speaker Recognition // Proc. Workshop on Speaker Recognition and its Commercial and Forensic Applications. Avignon, France, 1998. P. 211—214.

Сведения об авторах

Сергей Александрович Новосёлов — канд. техн. наук; ООО „ЦРТ-инновации", Санкт-Петербург; научный сотрудник; E-mail: novoselov@speechpro.com

Владислав Александрович Сухмель — аспирант; Санкт-Петербургский государственный университет, кафедра компьютерного моделирования и многопроцессорных систем; E-mail: sukhmel@apmath.spbu.ru

Алексей Владимирович Шолохов — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: sholokhovalexey@gmail.com

Тимур Сахиевич Пеховский — канд. физ.-мат. наук; ООО „ЦРТ-инновации“, Санкт-Петербург; ведущий научный сотрудник; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: tim@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

УДК 621.391.037.372

В. Л. Щемелинин, К. К. Симончик

ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ ГОЛОСОВОЙ ВЕРИФИКАЦИИ К АТАКАМ, ИСПОЛЬЗУЮЩИМ СИСТЕМУ СИНТЕЗА

Проанализирована устойчивость современных методов верификации к взлому при помощи гибридной системы синтеза речи на основе технологий Unit Selection и скрытых марковских моделей. Представлен метод взлома, обеспечивающий достижение ошибки ложного пропуска в 98—100 % случаев при большом объеме обучающей базы; метод может быть автоматизирован при сопряжении с автоматической системой распознавания речи.

Ключевые слова: *спуфинг, синтез речи, распознавание диктора.*

Введение. Системы верификации дикторов по голосу широко используются в криминалистических экспертизах, системах контроля доступа, банковской сфере, а также Интернете. Основные задачи подобных систем — повышение удобства использования и защита от несанкционированного доступа [1]. Соревнования NIST SRE 2012 [2] показали, что преобладают системы, основанные на представлении модели голоса диктора в пространстве полной изменчивости (*total variability*). Однако, как показывают исследования, современные системы верификации неустойчивы к спуфингу [3] с помощью автоматического синтеза голоса.

В настоящей работе исследована зависимость надежности системы верификации от объема речевого материала для обучения системы синтеза.

Система голосовой верификации. Предлагаемый метод заключается в использовании смесей гауссовых распределений (Gaussian Mixture Models, GMM) для моделирования голоса диктора, а затем их редукции до так называемого *i*-вектора в низкоразмерном пространстве полной изменчивости.

В работе использованы система текстозависимой верификации дикторов на базе *i*-векторов [4, 5], а также специальный модуль препроцессинга, включающий энергетический детектор речи и детектор клиппированных сигналов [6] для их отбраковки. В качестве речевых признаков выступали векторы мел-частотных кепстральных коэффициентов (Mel-frequency Cepstrum Coefficients, MFCC), их производных первого и второго порядка (39 элементов). Длина каждого речевого кадра для вычисления MFCC составляла 22 мс со сдвигом 11 мс. Для компенсации эффекта Гиббса использовалось взвешивание сигнала окном Хем-

минга. Эффекты канальных искажений на уровне признаков компенсировались путем вычитания кепстрального среднего (*cepstral mean subtraction*). Выравнивание признаков (*feature warping*) [7] не применялось, поскольку длительность речевого сигнала была мала.

На этапе моделирования голоса диктора использовалась гендеронезависимая универсальная фоновая модель (Universal Background Model, UBM), представленная 512-компонентной смесью гауссовых распределений. Обучение UBM производилось с помощью стандартного EM-алгоритма на телефонной части речевых баз данных NIST SRE 1988—2010 [8, 9]. Для ускорения вычислений использовалась диагональная ковариационная матрица UBM. Общее число дикторов в обучающих базах данных — около 4000. Модуль оценки *i*-вектора (как и модуль линейного дискриминантного анализа) был обучен более чем на 60 000 телефонных и микрофонных записях из тех же речевых баз данных.

Модель GMM в низкоразмерном пространстве полной изменчивости представляется следующим выражением:

$$\mu = m + T\omega + \varepsilon,$$

где μ — супервектор параметров GMM модели диктора, m — супервектор параметров UBM, T — матрица, задающая базис в редуцированном пространстве признаков, ω — *i*-вектор в редуцированном пространстве признаков, $\omega \in N(0,1)$, ε — вектор ошибки.

Метод взлома системы верификации. Известны различные способы спуфинга (от англ. *spoofing* — получение доступа обманным путем). Например, в работе [10] описываются способы на основе воспроизведения записи голоса, манипуляции с записью голоса, прикрытия рта носовым платком или закрытия носа рукой. В настоящей работе для спуфинга используется гибридный метод синтеза на основе Unit Selection и скрытых марковских моделей (Hidden Markov Model, HMM) [11].

Метод взлома предполагает создание синтезированного голоса пользователя системы верификации. Для обучения системы синтеза используется предварительно записанная спонтанная речь пользователя. На этапе текстозависимой верификации при помощи синтезированного голоса и перехваченного парольного текста создается парольная фраза, используемая далее для попытки верификации. Детальная схема атаки представлена на рис. 1.

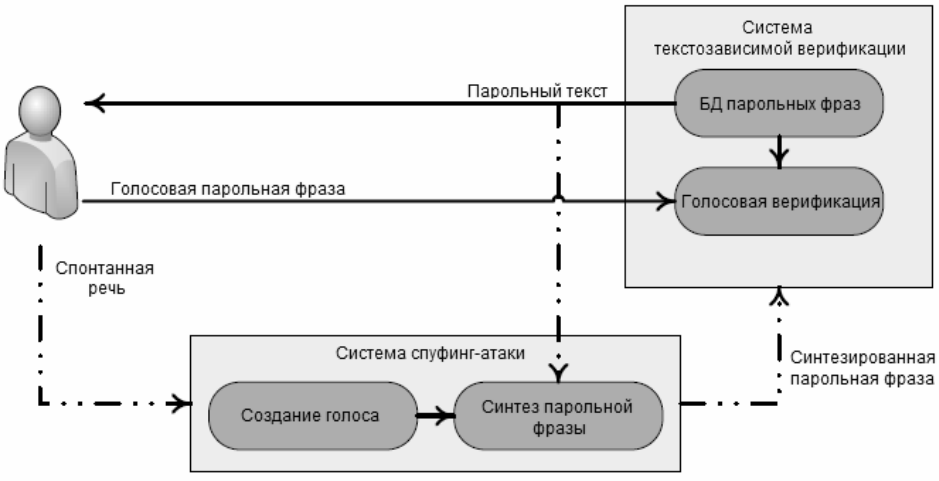


Рис. 1

Система синтеза голоса. Для моделирования спуфинг-атаки была использована система голосового синтеза, разработанная в ООО „ЦРТ“ [12]. В ней использованы два наиболее популярных подхода:

1) алгоритм Unit Selection (выбор речевых элементов), позволяющий достичь максимальной естественности синтезированной речи, при условии корректно отсегментированной на разных уровнях сбалансированной речевой базы данных большого объема;

2) статистические модели (HMM-синтез) позволяют легко модифицировать характеристики голоса с помощью адаптации/интерполяции дикторов. Речь, полученная на основе HMM-технологии, на слух менее естественна, однако в ней отсутствуют резкие, не обусловленные контекстом перепады по частоте и энергии, обычно присущие конкатенативному синтезу. Кроме того, применение HMM-синтеза позволяет разрабатывать новый голос за гораздо меньшее время, а также требует значительно меньше памяти для хранения речевой базы.

Речевой корпус. Для проведения экспериментов была использована речевая база русского языка, содержащая 7 различных дикторов (2 мужчин и 5 женщин), голоса которых использовались для обучения синтеза речи. Для каждого диктора было записано по 9 парольных фраз (2—3 секунды речи). Примеры парольной фразы: „Город Екатеринбург, улица вокзальная, дом 22, вокзал“, „Заплатить три рубля и дать объявление в бюллетене“ и т.п. Важно отметить, что записанные фразы не использовались в дальнейшем при обучении системы синтеза речи. Итого было записано 63 фразы различных дикторов.

Влияние обучающих данных системы синтеза на надежность верификации. Цель экспериментов — установить зависимость ошибки ложного принятия верификации (FA) от длительности речевого фрагмента, используемого при обучении системы синтеза голоса. Для экспериментов была взята описанная ранее система верификации по голосу. Калибровка порогов срабатывания системы производилась на речевой базе YOHO [13], содержащей 138 дикторов (мужчины и женщины), каждый из которых произносил фиксированную парольную фразу вида „36-24-36“ длительностью около (1,5—2 секунды активной речи).

Были определены два порога системы верификации:

1) по равновероятной ошибке пропуска—отклонения (equal error rate, EER) — ThresholdEER. На калибровочной базе EER=4 %;

2) при вероятности ложного принятия не более 1 % — ThresholdFA1. Этот порог обычно используется, когда необходимо обеспечить максимальную защиту системы от доступа злоумышленника.

Далее для каждого диктора выполнялись попытки доступа в систему верификации с помощью синтезированных парольных фраз, подготовленных системой синтеза голоса; объем спонтанной речи, использованной для обучения, — от 1 минуты до 4 часов речи для каждого диктора. На рис. 2 приведены результаты эксперимента, где *1* — калибровочный FA, *2* — 1 мин, *3* — 3 мин, *4* — 8 мин, *5* — 30 мин, *6* — 4 часа (*N* — мера сходства).

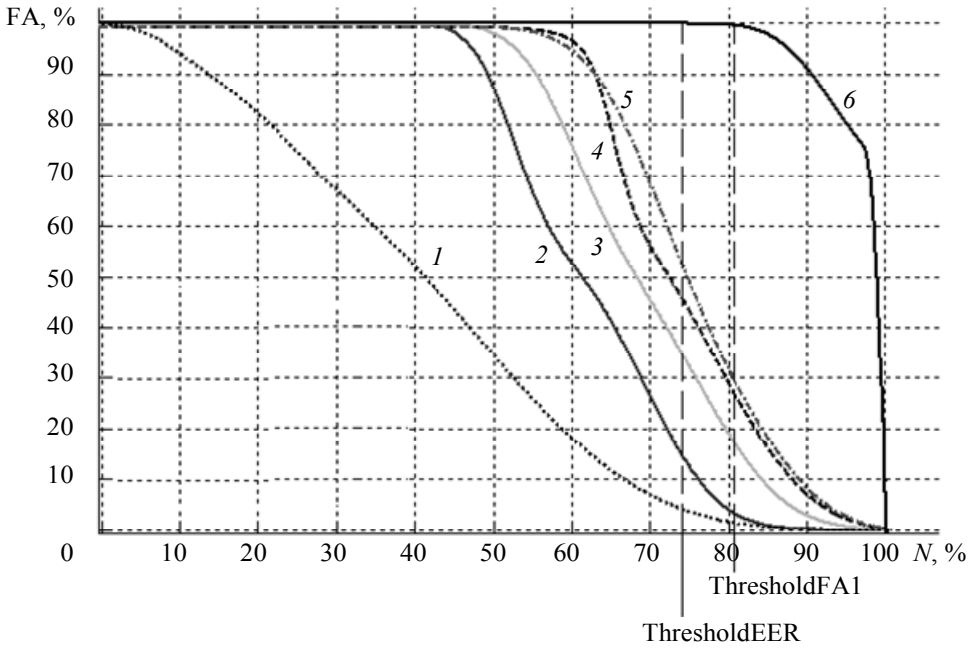


Рис. 2

В таблице представлены значения вероятности ложного принятия для двух порогов исследуемой системы верификации. Видно, что надежность системы верификации значительно снижается при использовании спонтанной речи длительностью от 8 минут и более; для системы верификации синтезированная речь практически перестает отличаться от живой речи человека при подготовке на данных большого объема (4 часа речи).

Объем речи для обучения синтеза	Ошибка FA (%) для порога	
	ThresholdEER	ThresholdFA1
1 минута	12,7 (8)	1,5 (1)
3 минуты	34,9 (22)	7,9 (5)
8 минут	44,4 (28)	19,1 (12)
30 минут	55,6 (35)	23,8 (15)
4 часа	100 (63)	98,4 (62)

На основании полученных результатов можно сделать выводы о том, что предлагаемый метод спуфинга позволяет не только серьезно ослаблять надежность системы верификации, но и обходить такую дополнительную меру защиты, как детектор присутствия диктора. Если система верификации передает пользователю пароль в виде звукового сообщения, возможно использование системы распознавания речи для полной автоматизации процесса спуфинга. В отличие от спуфинга путем конвертации признаков речи [14—16], предложенный подход, при использовании совместно с системой распознавания речи, позволяет исключить участие человека из диалога с системой верификации.

Выводы. В статье проанализирована устойчивость современных методов верификации к спуфингу при помощи гибридной системы синтеза речи на основе технологий Unit Selection и HMM. Как показали эксперименты, уже при использовании 8 и более минут обучающего материала, возможно существенно снизить надежность системы верификации, а при увеличении обучающего материала до четырех часов система практически не отличает синтезированный звук от речи диктора.

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

СПИСОК ЛИТЕРАТУРЫ

1. *Матвеев Ю. Н.* Технологии биометрической идентификации личности по голосу и другим модальностям // Вестн. МГТУ. „Приборостроение“. № 3 „Биометрические технологии“. 2012. С. 46—61.

2. The NIST Year 2012 Speaker Recognition Evaluation Plan [Electronic resource]: <http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf>.

3. *Wu Z., Kinnunen T., Chng E. S., Li H., Ambikairajah E.* A Study on spoofing attack in state-of-the-art speaker verification: the telephone speech case // Proc. of the APSIPA ASC 2012. Hollywood, USA, 2012. P. 1—5.

4. *Kenny P.* Bayesian speaker verification with heavy tailed priors // Proc. of the Odyssey Speaker and Language Recognition Workshop. Brno, Czech Republic, 2010.

5. *Simonchik K., Pekhovsky T., Shulipa A., Afanasyev A.* Supervized Mixture of PLDA Models for Cross-Channel Speaker Verification // Proc. of the 13th Annual Conf of the Intern. Speech Communication Association, Interspeech-2012. Portland, Oregon, USA, 2012. P. 1682—1685.

6. *Алейник С. В., Матвеев Ю. Н., Раев А. Н.* Метод оценки уровня клиппирования речевого сигнала // Научно-технический вестник информационных технологий, механики и оптики. 2012. № 3 (79). С. 79—83.

7. *Pelecanos J., Sridharan S.* Feature warping for robust speaker verication // Proc. Speaker Odyssey.The Speaker Recognition Workshop. Crete, Greece, 2001. P. 243—248.

8. *Matveev Yu. N., Simonchik K. K.* The speaker identification system for the NIST SRE 2010 // The 20th Intern. Conf. on Computer Graphics and Vision, GraphiCon’2010. St. Petersburg, 2010. P. 315—319.

9. *Козлов А. В., Кудашев О. Ю., Матвеев Ю. Н., Пеховский Т. С., Симончик К. К., Шулипа А. К.* Система идентификации дикторов по голосу для конкурса NIST SRE 2012 // Тр. СПИИРАН. 2013. Т. 25, № 2. С. 350—370.

10. *Villalba J., Lleida E.* Speaker verification performance degradation against spoofing and tampering attacks // Proc. FALA 10 Workshop. 2010. P. 131—134.

11. *Chistikov P. G., Korolkov E. A., Talanov A.O.* Combining HMM and unit selection tech nologies to increase naturalness of synthesized speech // Proc. of the Annual Intern. Conf. “Dialog-2013”. P. 2—10.

12. *Chistikov P. G., Korolkov E. A.* Data-driven Speech Parameter Generation For Russian TexttoSpeech System. Computational Linguistics and Intellectual Technologies // Proc. of the Annual Intern. Conf. ”Dialogue”. 2012. Vol. 1, Is. 11. P. 103—111.

13. *Campbell J., Higgins A.* YOHO Speaker Verification [Electronic resource]: <http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC94S16>.

14. *Wu Z., Chng E. S., Li H.* Speaker verification system against two different voice conversion techniques in spoofing attacks: Technical report [Electronic resource]: <http://www3.ntu>.

15. *Kinnunen T., Wu Z.-Z., Lee K. A., Sedlak F., Chng E. S., Li H.* Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech // Proc. ICASSP. Kyoto, Japan, 2012. P. 4401–4404.

16. *Aylett M. P., Yamagishi J.* Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. 2008.

Сведения об авторах

Вадим Леонидович Щемелинин — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: shchemelinin@speechpro.com

Константин Константинович Симончик — канд. техн. наук, доцент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела; E-mail: simonchik@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.13 г.

SUMMARY

P. 7—11.

ANALYSIS OF LANGUAGE STATISTICAL ASPECTS AND THEIR GENDER VARIATIONS BY THE EXAMPLE OF LITHUANIAN

Language aspects applicable for automatic language and speaker recognition are revealed. A method proposed for language recognition is based on statistical parameters of pitch stress pattern in a given language. Typical ranges of the parameters variation for languages of different language families are compared.

Keywords: speech technologies, statistical aspects of intonation, Lithuanian.

Data on authors

Mikhail V. Khitrov — Cand. Techn. Sci.; STC Ltd., St. Petersburg; General Director; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; Head of the Department; E-mail: khitrov@speechpro.com

Andrey Yu. Vasiliev — STC Ltd., St. Petersburg; Programmer; E-mail: vasilyev-a@speechpro.com

P. 12—17.

DETERMINATION OF CHANNEL-INDEPENDENT INFORMATION INDICATORS

Information indicators of speech are analyzed for creation of channel-independent feature space aimed at improvement of speaker recognition system efficiency. For the problem of determination of similarity between several audio recordings, the optimal set of channel-independent information feature vectors is determined experimentally with the use of dynamic time warping.

Keywords: speech analysis, machine learning, feature selection, mel-frequency cepstral coefficients, dynamic time warping.

Data on authors

Vitaly V. Kiselev — Speech Technologies Ltd., Minsk; Director; E-mail: kiselev-v@speechpro.com

Andrey V. Tkachenia — Speech Technologies Ltd., Minsk; Junior Researcher; E-mail: tkachenia-a@speechpro.com

Mikhail V. Khitrov — Cand. Techn. Sci.; STC Ltd., St. Petersburg; General Director; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; Head of the Department; E-mail: khitrov@speechpro.com

P. 17—22.

ANALYSIS OF DATA BALANCING PROBLEM IN ACOUSTIC MODELING OF AUTOMATIC SPEECH RECOGNITION SYSTEM

The problem of data balancing for training of acoustic models for automatic speech recognition system is considered. A metric is proposed which enables an explicit account for the data level in a cluster during triphone clustering. The proposed approach is shown to improve the quality of speech recognition.

Keywords: automatic speech recognition, GMM-HMM, acoustic modeling, acoustic model training, state tying, data balancing, model clustering, triphones.

Data on authors

Natalia A. Tomashenko — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC Ltd., St. Petersburg; Junior Researcher, E-mail: tomashenko-n@speechpro.com

Yury Yu. Khohlov — STC Ltd., St. Petersburg; Leading Programmer; E-mail: khokhlov@speechpro.com

P. 23—28.

CROSS-VALIDATION STATE CONTROL IN ACOUSTIC MODEL TRAINING OF AUTOMATIC SPEECH RECOGNITION SYSTEM

A technique is presented for optimization of Gaussian mixture models (GMM) size during the training of hidden Markov models (HMM), an essential part of many of the automatic speech recognition systems. Application of the technique increases recognition accuracy by avoiding the over-fitting effect, and reduces significantly computational load of the recognition procedure.

Keywords: automatic speech recognition, hidden Markov model, cross-validation control, cross-validation criteria.

Data on authors

German A. Chernykh — Cand. Phys.-Math. Sci.; St. Petersburg State University; STC Ltd., St. Petersburg; Researcher; E-mail: chernykh@speechpro.com

Maksim L. Korenevsky — Cand. Phys.-Math. Sci.; STC-Innovation Ltd., St. Petersburg; Researcher; E-mail: korenevsky@speechpro.com

Kirill E. Levin — Cand. Techn. Sci.; STC Ltd., St. Petersburg; Department of Speech Recognition; Head of the Department; E-mail: levin@speechpro.com

Irina A. Ponomareva — STC Ltd., St. Petersburg; Researcher; E-mail: ponomareva@speechpro.com

Natalia A. Tomashenko — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC Ltd., St. Petersburg; Junior Researcher, E-mail: tomashenko-n@speechpro.com

P. 28—32.

STATISTICAL METHODS FOR AUTOMATIC PROSODIC BREAK DETECTION IN A TEXT-TO-SPEECH SYSTEM

Application of statistical methods for predicting positions and durations of prosodic breaks in a text-to-speech system is proposed. The methods are shown to ensure better results as compared with a baseline rule-based system.

Keywords: prosodic breaks, prosodic boundaries, pauses, speech synthesis, statistical models.

Data on authors

Pavel G. Chistikov — STC Ltd., St. Petersburg; Researcher; E-mail: chistikov@speechpro.com

Olga G. Khomitsevich — PhD; STC Ltd., St. Petersburg; Senior Researcher; E-mail: khomitsevich@speechpro.com

Sergey V. Rybin — Cand. Phys.-Math. Sci.; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: rybin@speechpro.com

P. 33—39.

TIME DELAY ESTIMATION OF AUDIO SIGNALS USING THEIR ENVELOPES

The problem of time delay estimation for dual-channel acoustic signals (speech, music, etc.) recorded under reverberant conditions is investigated. A method of time delay estimation based on cross-correlation of temporal envelopes of the signals is presented. Comparison with other known methods of time delay estimation is provided.

Keywords: cross-correlation, delay estimation, signal envelope, signal processing.

Data on authors

Sergey V. Aleinik — STC-Innovation Ltd., St. Petersburg; Researcher; E-mail: aleinik@speechpro.com

Mikhail B. Stolbov — Cand. Techn. Sci.; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC-Innovation Ltd., St. Petersburg; Senior Researcher; E-mail: stolbov@speechpro.com

P. 40—47.

SPEECH SIGNALS STOCHASTICITY AND ITS EVALUATION

The known and new presented methods for evaluation of speech signal stochasticity are analyzed. Results of statistical simulation demonstrate advantages of the proposed approach as compared with the known ones: the estimates obtained with the new method possess lower variance and bias.

Keywords: stochasticity, speech signals, adaptive filtration, linear prediction.

Data on authors

Sergey V. Aleinik — STC-Innovation Ltd., St. Petersburg; Researcher; E-mail: aleinik@speechpro.com

Mikhail B. Stolbov — Cand. Techn. Sci.; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC-Innovation Ltd., St. Petersburg; Senior Researcher; E-mail: stolbov@speechpro.com

P. 47—52.

ASSESSMENT OF FUNCTIONAL SAFETY OF DETECTION OF VIBROACOUSTIC SIGNAL FROM ARRIVING TRAIN WITH ENERGY SENSOR

Functional safety of detection of a vibroacoustic signal from arriving train with energy detector is investigated. The lower value of detectability level is derived from a proposed false alarm probability. Sufficiency of developed method for detection of arriving train in the case of long-welded rails is demonstrated.

Keywords: arriving train, vibroacoustic signal, energy detector.

Data on authors

Sergey V. Bibikov — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC Ltd., St. Petersburg; Deputy Director; E-mail: bibikov@speechpro.com,

Yury N. Matveev — Dr. Techn. Sci., Professor; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC-Innovation Ltd., St. Petersburg; Chief Researcher; E-mail: matveev@mail.ifmo.ru

Nikolay N. Semenov — Cand. Techn. Sci.; STC Ltd., St. Petersburg; Team Manager; E-mail: semenov-n@speechpro.com

P. 53—57.

TARGET AND NON-TARGET SPEECH SEPARATION USING A DUAL MICROPHONE SYSTEM

A practical speech detection method for robust automatic speech recognition is proposed. The method employs a system of two symmetrical microphones oriented in opposite directions. The algorithm of signal processing allows for spatial filtering of speakers.

Keywords: speech activity detection, multi-channel audio, crosstalk.

Data on authors

Mikhail B. Stolbov — Cand. Techn. Sci.; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC-Innovation Ltd., St. Petersburg; Senior Researcher; E-mail: stolbov@speechpro.com

Marina Yu. Tatarnikova — STC-Innovation Ltd., St. Petersburg; Senior Researcher; E-mail: tatmar@speechpro.com

P. 58—63.

EXPERT SYSTEMS AND METHODS FOR SPEAKER IDENTIFICATION

Modern approaches to forensic phonographic examination are analyzed. Utilizing different Various software used for the purpose of speaker identification is considered. Special phonographic editor SIS II developed by the Speech Technology Center is described.

Keywords: forensic phonographic examination, speaker identification, software.

Data on authors

Elena V. Bulgakova — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: bulgakova@speechpro.com

Ekaterina V. Krasnova — STC Ltd., St. Petersburg; Researcher; E-mail: krasnova@speechpro.com

P. 63—70.

CONCEPT OF THE NATIONAL VOICE ACCOUNTING AND VOICE BIOMETRIC SEARCH SYSTEM

Concept of the national voice accounting and voice biometric search system is presented.

Keywords: voice accounting, voice biometric search, system infrastructure.

Data on authors

Dmitry V. Dyrmovsky — St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC Ltd., St. Petersburg; Head of Branch Office; E-mail: ddv@speechpro.com

Sergey L. Koval — Cand. Techn. Sci.; STC Ltd., St. Petersburg; Chief Expert; E-mail: koval@speechpro.com

Mikhail V. Khitrov — Cand. Techn. Sci.; STC Ltd., St. Petersburg; General Director; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; Head of the Department; E-mail: khitrov@speechpro.com

P. 70—76.

ANALYSIS OF MANIFOLD LEARNING METHODS APPLICABILITY TO SPEAKER RECOGNITION

Applicability of manifold learning methods widely used in image recognition, to the problem of speaker identification, is considered. Experimental study are carried out, the results are analyzed.

Keywords: manifold learning, speaker recognition.

Data on authors

Yury N. Matveev — Dr. Techn. Sci., Professor; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC-Innovation Ltd., St. Petersburg; Chief Researcher; E-mail: matveev@mail.ifmo.ru

Andrey K. Shulipa — STC-Innovation Ltd., St. Petersburg; Senior Researcher; E-mail: shulipa@speechpro.com

P. 77—84.

EMPLOYMENT OF DTW-BASED HMM-GMM MULTI-SESSION TRAINING IN TEXT-DEPENDENT SPEAKER VERIFICATION

An HMM training procedure using several password utterances is proposed. The proposed method is based on the Dynamic Time Warping algorithm, and is shown to allow for reduction of verification system errors.

Keywords: text-dependent speaker verification, short utterance, MFCC, HMM, GMM, DTW.

Data on authors

Sergey A. Novoselov — Cand. Techn. Sci.; STC-Innovation Ltd., St. Petersburg; Researcher; E-mail: novoselov@speechpro.com

Vladislav A. Sukhmel — Post-Graduate Student; St. Petersburg State University, Department of Computer Modelling and Multiprocessor Systems; E-mail: sukhmel@apmath.spbu.ru

Alexey V. Sholokhov — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: sholokhovalexey@gmail.com

Timur S. Pekhovsky — Cand. Phys.-Math. Sci.; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC-Innovation Ltd., St. Petersburg; Senior Researcher; E-mail: tim@speechpro.com

P. 84—88.

STUDY OF VOICE VERIFICATION SYSTEM TOLERANCE TO SPOOFING ATTACKS USING A TEXT-TO-SPEECH SYSTEM

A method of spoofing text-dependent voice verification system based on the most popular TTS approaches (Unit Selection and HMM) is presented. The method is shown to allow for false acceptance error of 98—100 % in the case of sufficiently large TTS database. A distinctive feature of the method is that it can be fully automated if used in conjunction with a speech recognition system.

Keywords: spoofing, speech synthesis, speaker recognition.

Data on authors

Vadim L. Shchemelinin — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: shchemelinin@speechpro.com

Konstantin K. Simonchik — Cand. Techn. Sci.; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; STC Ltd., St. Petersburg; Head of Department; E-mail: simonchik@speechpro.com