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 56 **ФЕВРАЛЬ 2013** **№ 2**

ТЕМАТИЧЕСКИЙ ВЫПУСК

РЕЧЕВЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

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

СОДЕРЖАНИЕ

THEMATIC ISSUE

SPEECH INFORMATION SYSTEMS

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

CONTENTS

Editor-in-Chief E. B. Yakovlev

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
Department Researcher, STC-Innovation Ltd.
Dr. Techn Sci.
Yu. N. MATVEEV

МЕТОДИЧЕСКИЕ И АЛГОРИТМИЧЕСКИЕ ОСНОВЫ ОБРАБОТКИ И АНАЛИЗА РЕЧЕВЫХ И ЗВУКОВЫХ СИГНАЛОВ

УДК 811.161.1

Н. С. Смирнова, М. В. Хитров

ФОНЕТИЧЕСКИ ПРЕДСТАВИТЕЛЬНЫЙ ТЕКСТ ДЛЯ ФУНДАМЕНТАЛЬНЫХ И ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ РУССКОЙ РЕЧИ

Приведен фонетически представительный текст, разработанный с применением новейших достижений в области лингвистических технологий. Полнота покрытия текстом фонетических единиц русской речи позволяет использовать его при формировании речевых корпусов для разработки и оценки экспертных и автоматических речевых систем различного назначения.

Ключевые слова: *фонетически представительный текст, фонетически сбалансированный текст, статистические характеристики русской речи, частотность и дистрибуция фонетических единиц.*

Если обратиться к прикладным областям речевых исследований, то можно заметить, что сегодня использование небольших фонетически представительных текстов при создании автоматических систем синтеза и распознавания речи уже не столь актуально, и приоритет отдается машинным методам статистического моделирования с использованием обучающих массивов текстовых и речевых данных очень больших объемов. Такие массивы данных называют также базами данных или корпусами (текстовыми или речевыми). Часто под корпусом понимают преимущественно те массивы данных, которые переведены в электронную форму и специальным образом обработаны, структурированы и аннотированы для целей разработки речевых приложений [1]. В настоящей статье опорным текстовым корпусом будет называться большой по объему массив текстов различного жанра, использованный нами для получения опорного статистического распределения фонетических единиц русской речи.

Однако наряду с разработкой речевых систем не менее актуальной остается задача выработки объективных критериев оценки их качества, и в этом случае тестовым материалом для оценки и сравнения систем автоматического синтеза и распознавания речи могут стать небольшие фонетически представительные тексты (ФПТ), позволяющие оценить полноту покрытия системой фонетических единиц целевого языка и выявить возможные недостатки ее работы. Кроме того, на таких текстах удобно проводить быструю подстройку системы под нового диктора.

В общем случае под фонетически представительным (репрезентативным) понимается такой текстовый материал, в котором частотное распределение фонетических единиц (фонем, аллофонов, слогов) соответствует общеязыковому распределению, получаемому из статистического анализа опорного текстового корпуса. В задачах, предполагающих исследование

региональной вариативности речевых характеристик, в качестве дополнительного критерия фонетической представительности текста должно рассматриваться наличие фонетических позиций и контекстов, способствующих проявлению региональной речевой специфики говорящего.

Фонетическая представительность, подобно фонетической сбалансированности, естественным образом предполагает присутствие в тексте всех фонем целевого языка в их основной дистрибуции. Фонетически сбалансированные и фонетически представительные тексты традиционно используются в качестве материала для изучения фонетических характеристик звучащей речи. Преимущество использования фонетически представительных текстов состоит, прежде всего, в их компактности наряду с информационной насыщенностью. С одной стороны, такие тексты обычно невелики по объему, а с другой — отражают фонетическое многообразие языковой системы не хуже произвольно взятых текстовых массивов значительного объема. Это достигается путем кропотливой работы по конструированию текста — наполнением его словами, содержащими требуемые фонетические единицы, а также сокращением его объема путем удаления элементов с низкой информативностью. В результате получается удобный для прочтения материал (обычно не более 600 слов), позволяющий исследовать характер реализации и варьирования в речи носителей определенного языка значимых фонетических характеристик и сформировать полноценный речевой портрет говорящего.

В русистике известно несколько фонетически представительных текстов, составленных на основе списков наиболее частотных слогов, приведенных в работах В. М. Елкиной и Л. С. Юдиной [2, 3]. Один из них — „Был тихий серый вечер“ [4] — лег в основу материала для Фонетического фонда русского языка.

Слог традиционно считается минимальной произносительной единицей, и потому оценка встречаемости слогов может лечь в основу формирования текстового материала. Однако на частоту встречаемости и состав выделяемых слогов оказывает влияние ряд факторов, в частности, характер опорного текстового материала (на основе которого получены показатели частотности слогов), используемая система транскрипции текста и степень ее подробности, а также принятая стратегия слогоделения. В отношении частотного распределения слогов, приведенного в работах [2, 3] и использованного впоследствии при составлении текста „Был тихий серый вечер“, следует отметить, что оно было получено на текстах радиотехнической тематики и с применением довольно спорной теории деления на открытые слоги, предложенной Л. В. Бондарко [5]. Кроме того, в классификации [2, 3] не различаются предударные и заударные слоги (а для некоторых гласных фонем — также ударные и безударные варианты), что приводит к серьезным упрощениям в оценках частотности и сочетаемости аллофонов русских фонем.

Исследовав методологию и инструментарии, ранее использовавшиеся при составлении фонетически представительных текстов, авторы разработали несколько иной подход, предполагающий, в частности, применение более подробной транскрипции текстового материала (с учетом предударной/заударной позиции гласного), преимущественно стилистически нейтрального текстового материала для получения опорной статистики, а также увеличение объема опорного текстового материала. Кроме того, поскольку существующие теории слогоделения допускают вариативность межслоговых границ для сочетаний ГС и СС (Г — гласный, С — согласный) и, как следствие, по-разному представляют состав и количество слогов русского языка (ср., например, принципы, предложенные М. В. Ломоносовым, Р. И. Аванесовым, Л. В. Щербой, Л. В. Бондарко [6]), было решено в качестве базовых единиц при составлении текста использовать последовательности СГ, поскольку при любом подходе они относятся к одному слогу. При этом в последовательностях типа ГГ в качестве самостоятельных элементов выделялись гласные, а на конце слова допускались закрытые слоги типа СГС. В качестве дополнительных критериев учитывалась встречаемость двухфонемных и трехфо-

немных сочетаний (так называемых дифонов и трифонов). Кроме наиболее частотных фонетических единиц в текст были введены звуковые последовательности и позиции, диагностически важные для выявления региональной речевой специфики.

Данный принцип построения текста был предпочтен „слоговому“ как более адекватный и экономичный для получения фонетической представительности. Если следовать слоговому принципу построения текста, то для полноценного выявления региональных особенностей говорящих потребовалось бы дополнительно включить в текст целый ряд низкочастотных звукосочетаний и позиций, что в комбинации с обеспечением высокой слоговой представительности неизбежно привело бы к увеличению объема текстового материала. Так, в составе частотных слогов отсутствует целый ряд элементов, чрезвычайно важных для исследования вариативности русской речи — в частности, конечный мягкий <-вь> и другие мягкие губные. Например, первый в списке по теории Аванесова [7] слог с конечным „ф’“ [к $а_{з/уд}$ ф’] имеет ранг 989, слог [б $О_{уд}$ ф’] — 1352, а первый по частотности слог с конечным мягким „п’“ (слово „степь“) — лишь 3993. Подобная ситуация наблюдается и в отношении ряда других важных в диагностическом плане звуковых элементов. Кроме того, известно, что на качественные характеристики гласных в русском языке преимущественное влияние оказывает левый контекст, и неслучайно при различных подходах к слогоделению именно последовательность СГ неизменно относится к одному слогу.

Материал для получения опорной статистики был скомпонован из текстов классической и современной литературы, а также современной публицистики (отекстованные интервью, репортажи, дискуссии). Он включает в себя более 460 тыс. словоформ, более 1 млн слогов (по сравнению с более 100 тыс. в работах [2, 3]), более 2,5 млн фонемоупотреблений. Была оценена встречаемость фонем (монофонов), звукосочетаний (двух- и трехфонемных) и слогов (по трем различным сценариям слогоделения). Кроме того, для ряда фонем был составлен список фонетических позиций и контекстов, потенциально значимых для выявления региональной вариативности русской речи (например, мягкие губные в конечной позиции, определенные сочетания согласных). Опорный текстовый корпус и его статистические характеристики приведены в работах [8, 9].

На основе статистик, полученных на опорном материале, с учетом фонемных позиций и комбинаций, способствующих выявлению региональной произносительной специфики, был составлен новый фонетически представительный текст. Он состоит из 533 слов, 1197 слогов (по числу гласных). Всего текст насчитывает 2902 фонемоупотребления. Текст включает в себя как описательную, так и богатую диалоговую часть (все коммуникативные типы); в нем представлены все фонемы русского языка во всех допустимых аллофонах (включая межсловные озвонченные аллофоны непарных глухих русских фонем / х /, / ч /, / ц / и / щ / — соответственно $х_{озв}$, $ч_{озв}$, $ц_{озв}$ и $щ_{озв}$). В тексте присутствует более 99 % сочетаний типа СГ (из них 98 % — наиболее частотные 258), 92 % возможных в русском языке двухфонемных сочетаний (из них 62 % — 250 наиболее частотных), значительно расширен (по сравнению с существующими текстами) набор сочетаний „согласный + ударный гласный“, возможных в русской речи. Для трех рассмотренных вариантов слогоделения (по [6, 7, 10]) доля покрытия типов слогов не ниже 70 % (что на 4—5 % выше, чем в тексте „Был … вечер“). Отметим также, что в текст целенаправленно были введены слова со звукосочетаниями, важными для исследования региональной и индивидуальной произносительный вариативности, в том числе иностранного происхождения, что привело к повышению доли низкочастотных слогов.

В табл. 1 приведены данные о встречаемости в нашем тексте аллофонов русских фонем в сопоставлении с их статистическим распределением в опорном текстовом корпусе. При обозначении аллофонов русских фонем используются следующие конкретизаторы: п/уд — предударный, з/уд — заударный, б/уд — безударный, озв — озвонченный, ’ — мягкий.

В тексте присутствует 56 типов аллофонов русских фонем (как уже упоминалось выше, отсутствуют лишь редкие безударные аллофоны фонемы / е /).

Таблица 1

Аллофон	Ранг в опорном корпусе	Ранг и встречаемость в тексте	Аллофон	Ранг в опорном корпусе	Ранг и встречаемость в тексте
$а_{1\text{-й п/уд}}$	1	1 (137)	т'	23	29 (39)
$а_{з/уд}$	2	2 (132)	ф	36	30 (38)
$и_{з/уд}$	3	3 (131)	$Ы_{уд}$	40	31 (38)
й	7	4 (124)	г	32	32 (36)
$и_{п/уд}$	4	5 (123)	в'	37	33 (34)
$О_{уд}$	9	6 (119)	ж	38	34 (34)
т	6	8 (115)	м'	39	35 (34)
$А_{уд}$	5	7 (114)	ч	28	36 (33)
н	8	9 (107)	$у_{п/уд}$	31	37 (32)
р	12	10 (101)	б	29	38 (31)
к	14	11 (95)	ш	30	39 (30)
с	10	12 (88)	$ы_{п/уд}$	43	40 (29)
в	11	13 (88)	ц	42	41 (27)
$Е_{уд}$	13	14 (86)	д'	35	42 (23)
м	17	15 (77)	х	41	43 (20)
д	21	16 (68)	к'	44	44 (17)
п	16	17 (63)	п'	45	45 (16)
л’	18	18 (62)	щ	46	46 (12)
$а_{2\text{-й п/уд}}$	20	19 (60)	б'	47	47 (11)
$ы_{з/уд}$	24	20 (54)	з'	49	48 (9)
н'	15	21 (54)	ф'	51	49 (7)
с'	25	22 (53)	$о_{б/уд}$	48	50 (5)
л	19	23 (51)	г'	50	51 (5)
$У_{уд}$	34	24 (50)	х'	53	52 (4)
р'	26	25 (47)	$х_{озв}$	52	53 (3)
з	27	26 (47)	$ч_{озв}$	55	54 (1)
$у_{з/уд}$	33	27 (44)	$ц_{озв}$	56	55 (1)
$И_{уд}$	22	28 (42)	$щ_{озв}$	58	56 (1)

Как видно из табл. 1, распределение частотности аллофонов в разработанном тестовом материале достаточно близко к распределению в опорном корпусе.

Совпадает состав 14 наиболее частотных аллофонов и 13 наиболее редких (разница в ранге — не более 3). В частотах остальных 39 аллофонов наблюдаются более существенные различия в рангах. В среднем разница в рангах составляет 2,89; максимальная разница в ранге наблюдается для ударного „У“ — 10.

Основные статистические характеристики созданного фонетически представительного текста приведены в табл. 2.

Таблица 2

Типы единиц	Типы единиц в ФПТ относительно опорного корпуса, %	Общее покрытие единиц опорного корпуса в ФПТ, %
Фонемы	96,6	99,9
Последовательности СГ	72,3	99,9
Дифоны	46,2	91,6
Трифоны	6,6	42,3
Слоги (по Аванесову)	6,2	74,1
Слоги (по Щербе)	5,9	72,6
„Открытые“ слоги	5,6	74,1

Из табл. 2 видно, что наш текст обеспечивает практически стопроцентное покрытие фонемного состава опорного текстового корпуса. Столь же высокий процент покрытия обеспечивают и присутствующие в сформированном тексте сочетания СГ (отсутствующие 118 типов таких последовательностей составляют менее 1 % опорного корпуса). Из числа возможных типов слога в разработанном тексте присутствует лишь 5—6 %, однако они покрывают 73—75 % всех слогов, встречающихся в опорном корпусе. В тексте встречается чуть менее половины (46 %) возможных в русском языке дифонов, однако при этом общая степень покрытия реализаций дифонов опорного корпуса достигает 92 %. Состав трифонов опорного корпуса наиболее обширен и насчитывает более 35 тыс. типов. Созданный текст включает более 2 тыс. типов трифонов (7 %), что покрывает 43 % всех реализаций трифонов опорного корпуса.

Всего в тексте присутствует 1197 гласных и 1705 согласных, консонантный коэффициент 1,42, что несколько выше, чем в опорной статистике (1,35 в опорном корпусе; 1,38—1,39 — по литературным источникам [11]). К более высокому значению консонантного коэффициента привело введение в текст слов с диагностическими консонантными последовательностями и позициями, а также дополнение текста словами с низкочастотными звуками и звукосочетаниями (в основном консонантными). Таким образом, фонемный состав текста был сбалансирован для получения более надежных результатов исследований.

Приведем фонетически представительный текст.

Дом, в котором я живу, расположен на окраине маленького городка, у самой подошвы горы. Здесь мягкий климат и редко идут дожди. Ночью небосвод бывает так густо усеян звездами, что кажется, будто все миллиарды их из нашей галактики разбросаны вверху над моей головой. Летним утром, как только я открываю окно, моя большая комната наполняется запахом цветов. Ветки черешен смотрят мне в окна, и легкий теплый ветер усыпает мой письменный стол белыми лепестками.

Я слушаю щебет птиц. Вот с искрометным задором пропел зяблик. Где-то дятел устраивает дупло. А это черные дрозды — поют не хуже соловьев. Прямо передо мной внизу — пестрый узор из крыш городских домов, а вдалеке, на краю горизонта, тянется серебряная цепь снеговых вершин... Весело жить в такой земле! Отрадное чувство разливается в жилах: вокруг величественные горы, воздух чистый и свежий, солнце яркое, небо синее — чего еще желать?.. Бьют настенные часы над камином: пять, шесть, семь, восемь, девять... Нужно торопиться в бюро. Минуты две-три ищу в шкафу электрическую схему, привезенную французским коллегой. Наконец заглядываю в портфель, нахожу ее внутри и вкладываю в книгу. После этого плотно закрываю жалюзи, однако сквозь щелки все равно пробивается солнечный свет. Выхожу на крыльцо и запираю ключом дверь.

Спустившись ниже к центру города, иду бульваром. Часть дороги проходит по пешеходному мосту через реку. Гибкие стальные тросы держат невысокий мост. Они привязаны к специальным тяжелым якорям, врытым в землю. Останавливаюсь в начале мостика у ограды, чтобы полюбоваться рельефными склонами горных хребтов, всматриваюсь в речную рябь. Под мостом с шумом плещутся мелкие рыбешки, возмущая водную гладь. Откуда-то доносится музыка: ноктюрн Шопена — позывные местной радиостанции.

Вдруг позади себя я слышу: „Сережа, неужели это ты? Вот так встреча!“. Я узнаю этот низкий голос. Оборачиваюсь — так и есть: Андрей Сафонов! Очень радостно видеть его вновь. С Андреем мы знакомы с конца восьмидесятых — служили вместе в армии. Его, энергичного и общительного, всегда на помощь готового прийти, любили все. Меня покорили его честность и недюжинная сила. Мы были дружны, но потом, мало-помалу, связь наша оборвалась.

— Здравствуй, Андрюша! Как ты тут оказался?

— Командировка в архив: предлагают снять сюжет про судоверфь. Вчера приехали — сегодня уезжаем. Вернее, улетаем — к четырем в аэропорт.

— Так скоро? И куда?

— Следующим пунктом Уфа. Прямой рейс. А ты, значит, теперь здесь живешь? Давно?

— Два года будет в феврале. Обменял свою городскую квартиру на бревенчатую избу.

— Серьезно? Не жалеешь?

В глазах моего приятеля мелькнул веселый огонек.

— Нисколько. Отдыхаю от километровых пробок, сутолоки и пыли.

— По-прежнему плывешь против течения? Счастливый ты человек, Сергей.

— Ладно, расскажи лучше о себе. Мы тысячу лет не виделись. Как жизнь? Как семья?

— Все у нас хорошо, все здоровы. Мы с женой работаем, дочь гимназию заканчивает.

— Ну а Федор как? Учится?

— Он в этом году поступил в медицинский.

— Какой молодец! Поздравляю вас!

— Спасибо.

— Кстати, ты позавтракал?

— Немного кофе выпил в гостинице. С удовольствием бы съел что-нибудь.

— Недалеко отсюда есть кафе. Мы привыкли там есть. Пойдем, провожу тебя. Вполне приличный сервис, разнообразное меню. Одно из их „фирменных“ блюд — рыба по-бенгальски. Рекомендую: вкус необыкновенный — для настоящих гурманов.

— Звучит слишком изысканно. Попроще ничего нет?

— Как насчет яичницы с грибами?

— В самый раз. А ты торопишься? Может быть, составишь мне компанию?

За разговорами и воспоминаниями незаметно пролетел завтрак. Приближалось время сказать „до свидания“. Мы расстались в твердом намерении больше не терять друг друга из виду.

СПИСОК ЛИТЕРАТУРЫ

1. *Кривнова О. Ф.* Фонетическое обеспечение для построения речевого корпуса // Акустика речи. Медицинская и биологическая акустика. Сб. тр. XIII сессии Российского акустического общества. Т. 3. М.: ГЕОС, 2003. С. 118—122.

2. *Елкина В. М., Юдина Л. С.* Статистика слогов русской речи // Вычислительные системы. Новосибирск, 1964. Вып. 10. С. 58—78.

3. *Елкина В. М., Юдина Л. С.* Статистика открытых слогов русской речи // Там же. Вып. 14. С. 55—91.

4. *Степанова С. Б.* Фонетические свойства русской речи: реализация и транскрипция: Дис. ...канд. филол. наук. Л., 1988.

5. *Бондарко Л. В.* Фонетика современного русского языка. Л., 1998. С. 196—211

6. *Бондарко Л. В.* Структура слога и характеристики фонем // Вопросы языкознания. 1967. № 1. С. 34—46.

7. *Аванесов Р. И.* О слогоразделе и строении слога в русском языке // Там же. 1954. № 6. С. 88.

8. *Смирнова Н. С., Чистиков П. Г.* Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задач в области речевых технологий // Матер. XXVII Междунар. конф. „Диалог“. М., 2011. С. 632—644.

9. *Smirnova N., Chistikov P.* Statistics of Russian Monophones and Diphones // Proc. of Specom-2011. Kazan, Russia, 2011. P. 218—223.

10. *Щерба Л.В.* Теория русского письма. Л., 1983. С. 29—33.

11. Фонетика спонтанной речи / Под ред. *Н. Д. Светозаровой.* Л., 1988. С. 210.

Сведения об авторах

Наталья Сергеевна Смирнова — канд. филол. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель группы лингвистов; E-mail: nsmirnova@speechpro.com

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 621.391.037.372

С. В. Алейник, М. Б. Столбов

ПОДАВЛЕНИЕ АКУСТИЧЕСКИХ ПОМЕХ АУДИОУСТРОЙСТВ С ИСПОЛЬЗОВАНИЕМ АСИНХРОННОГО ОПОРНОГО СИГНАЛА

Предложен метод двухканального шумоподавления для случая записи помехи, взятой из стороннего источника. Рассмотрены детали реализации разработанного метода, приведено сравнение его эффективности с эффективностью методов адаптивной компенсации помех.

Ключевые слова: *шумоподавление, акустические помехи, адаптивная обработка сигналов.*

Введение. Подавление помех в фонограммах является важной задачей для многих областей речевых технологий: идентификация диктора, восстановление старых фонограмм и т.п. Такая задача становится особенно актуальной, когда уровень помехи сопоставим с уровнем полезного речевого сигнала. Для ее решения предложено большое число различных алгоритмов шумоподавления [1—4].

Если помеха создается аудиоустройством и является нестационарной (пение, музыка и т.п.), эффективность одноканальных алгоритмов подавления шума уменьшается. В этом случае могут применяться двухканальные схемы адаптивной компенсации помех. В таких схемах сигнал в основном канале (основной сигнал) содержит смесь полезного речевого сигнала и помехи, а сигнал в опорном канале (опорный сигнал) содержит только помеху. Совместная обработка этих двух сигналов позволяет, при определенных условиях, эффективно подавлять помехи в основном сигнале.

Схема двухканального подавления помех представлена на рис. 1.

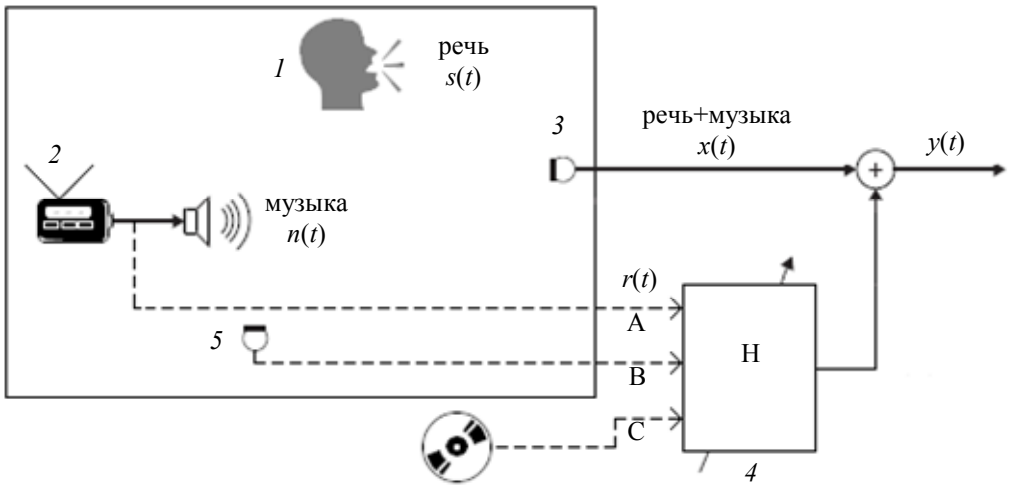


Рис. 1

Рассмотрим ситуацию записи фонограммы в помещении, когда речь $s(t)$ произносится человеком (*1*) на фоне акустической помехи $n(t)$, создаваемой работающим аудиоустройством (*2*). Речь и помеха принимаются микрофоном основного канала (*3*), формирующим основной сигнал $x(t)$. Целью обработки является подавление помехи и выделение речевого сигнала.

Процесс шумоподавления в двухканальных схемах (рис. 1) можно представить следующим образом. В дискретном случае сигналы основного и опорного каналов, $x(i)$ и $r(i)$ соответственно, описываются выражениями:

$$x(i) = h_{xs*}s(i) + h_{xn*}n(i),$$

$$r(i) = h_{rn*}n(i),$$

где i — временной индекс; $s(i)$ — речевой сигнал; $n(i)$ — помеха; * — символ свертки; h_{xs}, h_{xn} и h_{rn} — импульсные характеристики среды распространения для сигналов опорного и основного каналов.

Компенсация помехи (шумоподавление) в основном канале базируется на преобразовании опорного сигнала:

$$y(i) = x(i) + \mathrm{H}[r(i)],$$

где $y(i)$ — сигнал на выходе шумоподавителя; H — оператор преобразования опорного сигнала (рис. 1, блок *4*).

В зависимости от источника опорного сигнала возможна реализация различных алгоритмов обработки. В первом случае опорный сигнал снимается непосредственно с электрической цепи перед акустическим аудиоустройством (вход „A“ на рис. 1). В этом случае задача подавления помехи формулируется как задача эхоподавления [1, 2], в которой применяются алгоритмы адаптивной компенсации помех [2].

Во втором случае подавление помехи осуществляется с использованием опорного сигнала от микрофона, расположенного вблизи акустического источника помехи (микрофон *5*, вход „B“, рис. 1).

Наконец, особым случаем является ситуация, когда запись синхронного опорного сигнала отсутствует. Однако, если известно, какой звукоряд является помехой, то в качестве опорного сигнала может быть использована фонограмма, взятая из стороннего источника: например, музыкальная запись на компакт-диске (вход „C“ на рис. 1). В этом случае опорный сигнал является „асинхронным“, так как записан в другое время, на другой аппаратуре и в иных условиях [3].

Целью предлагаемой работы является описание практической реализации метода шумоподавления с использованием асинхронного опорного сигнала для случая акустических помех, создаваемых аудиоустройствами в помещениях.

Постановка задачи асинхронного шумоподавления. Эффективность шумоподавления зависит от целого ряда факторов: тип аудиосистемы, условия распространения звука в помещении, особенности практической реализации алгоритма шумоподавления и т.п. Из физических соображений ясно, что в асинхронном случае характеристики помех в опорном и основном каналах будут существенно различаться. Поэтому непосредственное использование асинхронной записи помехи оказывается неэффективным вследствие двух групп факторов.

1. Отсутствие синхронизации основного и опорного сигналов:

— несовпадение начала и конца помех в основном и опорном каналах;

— несовпадение частот дискретизации сигналов основного и опорного каналов.

2. Различие характеристик каналов записи сигналов основного и опорного каналов:

— различны условия записи музыки (например, запись оркестра на высококачественный CD и микрофонная запись сигнала тракта воспроизведения ТВ-приемника);

— записи выполнены в различных помещениях — различны характеристики среды (параметры реверберации и т.п.);

— различны частотные характеристики трактов записи.

Отсутствие синхронизации требует пояснения. Несовпадение начала помех в каналах связано с тем, что в асинхронном опорном сигнале помеха представляет собой полный звуко-

ряд, например, студийную запись музыкального произведения. Помеха в основном канале является только участком данного звукоряда, на который наложен полезный речевой сигнал. Начало данного участка может соответствовать любому месту музыкального произведения. Непростым является случай, когда короткий речевой сигнал начинается и заканчивается на участке, соответствующем припеву (или иному повторяющемуся фрагменту) в песне, что вызывает трудности в конкретной локализации участка помехи.

Несовпадение частот дискретизации опорного и основного сигналов также является общей проблемой для асинхронного случая. Обычно взятый с CD опорный сигнал представляет собой высококачественную запись, выполненную с частотой дискретизации 44 100 Гц. При этом основной сигнал дискретизирован с другой частотой, например, 11 025 Гц. В этом случае частота дискретизации опорного сигнала приводится к частоте основного с помощью известных алгоритмов. Однако даже после данной процедуры возможно незначительное различие в частотах дискретизации.

Такое различие приводит к тому, что в дискретизированных опорном и основном сигналах на одинаковый временной интервал приходится различное количество отсчетов. Например, в одном из случаев при анализе фонограмм частота дискретизации основного и опорного сигналов оказалась равна 16 и 16,0941 кГц соответственно, т.е. уже на десятой секунде разница в количестве отсчетов между опорным и основным сигналами составляла 941. Поскольку обработка велась покадрово, а размер кадра был выбран равным 512 отсчетам, то текущий и все последующие кадры уже не соответствовали друг другу, что привело к полной потере эффективности шумоподавления.

Различие условий записи основного и опорного сигналов в асинхронном случае также является важным фактором. Известно [2, 3], что эффективность шумоподавления адаптивных компенсаторов помех зависит от когерентности сигналов в опорном и основном каналах. Различие условий записи сигналов значительно снижает их когерентность, вследствие чего адаптивные компенсаторы оказываются малоэффективными.

Однако физические предпосылки для подавления шума в асинхронном случае все же существуют, поскольку помеха в основном и опорном каналах представляет собой различные реализации одного и того же звукоряда.

Для решения поставленной задачи нами был разработан полуавтоматический метод асинхронного шумоподавления, состоящий из двух основных шагов:

1) синхронизация основного и опорного сигналов;

2) подавление помехи в основном канале с использованием сигнала опорного канала.

Синхронизация основного и опорного сигналов представляет собой выполнение следующей последовательности действий:

— грубая синхронизация основного и опорного сигналов;

— точное совмещение начала помехи в основном и опорном сигналах;

— синхронизация частот дискретизации основного и опорного сигналов.

Грубая синхронизация выполняется оператором и включает в себя:

— приведение сигналов к единой частоте дискретизации (обычно это частота дискретизации сигнала основного канала);

— приведение средних спектров мощности сигналов к единому виду;

— приближенное определение (на слух, по спектрограмме и/или осциллограмме) начала и конца соответствующих друг другу участков помехи в опорном и основном каналах и расставление меток начала и конца участков помехи;

— приближенное совмещение участков начала помехи в опорном и основном сигналах.

Точное совмещение начала фрагментов с помехой в опорном и основном сигналах выполняется автоматически с использованием метода определения задержки сигнала по взаимокорреляционной функции [5]. Однако поскольку помехи в опорном и основном

каналах практически некоррелированы, то оценка по максимуму взаимокорреляционной функции сигналов неэффективна (максимум слабо выражен или отсутствует).

С другой стороны, кратковременные огибающие спектра мощности основного и опорного сигналов $P_x(t)$ и $P_r(t)$ на участках помехи оказываются в значительной степени коррелированными [6], так как кратковременные огибающие спектра мощности менее подвержены влиянию среды распространения и акустических трактов устройств записи—воспроизведения. Поэтому синхронизация осуществлялась по максимуму взаимной корреляции огибающих мощности опорного и основного сигналов $P_x(i)$ и $P_r(i)$:

$$P_x(i) = \left\langle x^2(i) \right\rangle \text{ и } P_r(i) = \left\langle r^2(i) \right\rangle,$$

где $\langle\ \rangle$ — символ сглаживания по времени; i — временной индекс.

С целью снижения временных затрат для оценки огибающих использовался алгоритм экспоненциального сглаживания:

$$P_x(i) = \alpha P_x(i-1) + (1-\alpha)x^2(i),$$

где $0 \le \alpha < 1$ — постоянная сглаживания, задаваемая таким образом, чтобы соответствовать темпу музыки, т.е. чтобы сигнал усреднялся без потери информации о колебаниях огибающей.

Далее, на начальных участках помехи в основном и опорном каналах (5—10 с) вычисляется взаимокорреляционная функция огибающих мощности $C(m)$:

$$C(m) = \sum_i (P_x(i) - \overline{P_x})(P_r(i-m) - \overline{P_r}),$$

где $\overline{P_x}$ и $\overline{P_r}$ — средние значения для $P_x(i)$ и $P_r(i)$ соответственно.

После этого для синхронизации начала помехи осуществляется сдвиг опорного сигнала на число отсчетов, соответствующих максимуму функции $C(m)$.

Точная синхронизация частот дискретизации также выполняется по максимуму взаимной корреляционной функции $C(m)$, вычисленной на участках, помеченных как окончание помехи в опорном и основном каналах. Если максимум $C(m)$ не соответствует нулевому сдвигу, то частоты дискретизации основного и опорного сигналов различаются. Тогда вычисляется относительный коэффициент сжатия/растяжения опорного сигнала:

$$S = (N_r + \arg\max\left(C(m)\right)\big/N_r,$$

где N_r — число отсчетов между метками начала и конца помехи в опорном сигнале. Если $S > 1$, то выполняется сжатие опорного сигнала, если $S < 1$ — то растяжение. Сжатие (растяжение) в экспериментах выполнялось на основе поотсчетной интерполяции, при этом линейная и квадратичная интерполяция давала практически одинаковые результаты.

Шумоподавление на основе метода спектрального вычитания. Опорный сигнал, полученный в результате точной синхронизации, может быть использован для компенсации помехи в основном канале. Однако применение линейных адаптивных компенсаторов в данном случае оказалось малоэффективным, что объясняется существенным уменьшением когерентности помехи в основном и опорном каналах вследствие различия условий записи и проведенных преобразований. Для этих условий наиболее подходит использование алгоритмов спектрального вычитания (АСВ) [7—9], поскольку АСВ не учитывают фазовых соотношений и позволяют подавлять помехи в случае их слабой когерентности в опорном и основном каналах.

Двухканальный АСВ организован следующим образом [1]. Мгновенный спектр Фурье на кадре основного сигнала может быть представлен в виде суммы спектров полезного речевого сигнала и спектра помехи:

$$X(f,k) = S(f,k) + N(f,k),$$

где f — частота и k — временной индекс кадра.

Спектральное вычитание определяется как [1]:

$$\left|Y(f,k)\right| = \left|X(f,k)\right| - \bar{N}(f,k),$$

где $Y(f,k)$ — оценка спектра выходного сигнала; $\bar{N}(f,k)$ — оценка амплитудного спектра помехи.

В этом случае $\left|Y(f,k)\right|$ может быть записан как:

$$\left|Y(f,k)\right| = G(f,k)\left|X(f,k)\right|,$$

где $G(f,k)$ — целевая функция фильтра шумоподавления вида:

$$G(f,k) = 1 - \bar{N}(f,k)/\left|X(f,k)\right|.$$

В более общем виде целевая функция определяется как [1]:

$$G(f,k) = \max\left[b, 1 - a\bar{N}(f,k)/\left|X(f,k)\right|\right],$$

где a и b — параметры алгоритма „коэффициент вычитания“ и „глубина подавления шума“ соответственно.

Спектр сигнала после шумоподавления рассчитывается с применением целевой функции фильтра к исходному комплексному спектру сигнала:

$$Y(f,k) = G(f,k)X(f,k).$$

Временной сигнал $y(i)$ на выходе шумоподавителя вычисляется путем обратного преобразования Фурье последовательности спектров $Y(f,k)$.

Поскольку спектр мощности шума в основном канале неизвестен, то в вычислениях используется его оценка, определяемая следующим образом. В реверберирующем помещении оценка комплексного спектра помехи может быть представлена как сумма спектров ранней и поздней реверберации [7]:

$$N(f,k) = A_0(f)R_a(f,k) + \sum_m A_m(f)R_a(f,k-m),$$

где $A_0(f)$ — фильтр, описывающий эффекты ранней реверберации; $A_m(f)$ — передаточные функции, соответствующие задержке на m кадров; $R_a(f,k)$ — комплексные спектры помехи.

Предполагая, что фазы спектров для отдельных кадров некоррелированы, мгновенный спектр мощности помехи аппроксимируем как:

$$\left|N(f,k)\right|^2 = \left|A_0(f)\right|^2\left|R_a(f,k)\right|^2 + \sum_m \left|A_m(f)\right|^2\left|R_a(f,k-m)\right|^2.$$

В рамках предлагаемого алгоритма нами учитывался только шум, порожденный ранней реверберацией, т.е.

$$N(f,k) = A_0(f)R_a(f,k).$$

В случае использования фонограммы в качестве опорного сигнала мгновенные спектры опорного сигнала $R(f,k)$ преобразуются в спектры акустической помехи путем умножения на частотный отклик $B(f)$ аудиосистемы:

$$R_a(f,k) = B(f)R(f,k).$$

Тогда спектр помехи в основном канале может быть представлен следующим соотношением:

$$\bar{N}(f,k) = A_0(f)B(f)R(f,k) = W(f,k)R(f,k),$$

где $W(f,k)$ — передаточная функция преобразования опорного сигнала. Передаточная функция может изменяться в зависимости от положения диктора и акустической обстановки в помещении, поэтому необходим адаптивный алгоритм ее оценки. В работе [9] предложен алгоритм адаптивной оценки передаточной функции в моменты присутствия акустической помехи в опорном канале и отсутствия речи диктора в основном канале. „Музыкальная“ помеха, как правило, присутствует непрерывно. При этом детектировать паузы в речи диктора представляется затруднительным ввиду нестационарного характера помехи, особенно при ее высоком уровне. Для подобного случая нами предложен следующий алгоритм оценки передаточной функции $\hat{W}(f,k)$:

$$\hat{W}(f,k) = \hat{W}(f,k-1) + \mu\left(|X(f,k)| - \hat{W}(f,k-1)|R(f,k)|\right) \Big/ \left(|X(f,k)|^2 + |R(f,k)|^2\right),$$

где $\mu < 1$ — скорость адаптации.

Экспериментальные исследования подтвердили работоспособность алгоритма оценки передаточной функции на разных типах тестовых и модельных сигналов.

С учетом оценки передаточной функции результирующий АСВ описывается следующим выражением:

$$G(f,k) = \max[b,\ 1 - a\hat{W}(f,k)|R(f,k)|/|X(f,k)|].$$

Экспериментальная оценка эффективности разработанного алгоритма. Работоспособность предложенного метода проиллюстрируем результатами следующего эксперимента. В помещении (6×5×3 м, время реверберации 480 мс) располагалась акустическая колонка. Через колонку проигрывались записанные в компьютере с частотой 16 кГц тестовые моносигналы длительностью 1,5 мин каждый: музыка, речь и розовый шум. Принятый через микрофон акустический сигнал основного канала записывался на цифровой диктофон. Микрофон основного канала в первой сессии располагался на расстоянии 1 м от акустической колонки; в последующих сессиях — на расстоянии d=2, 3 и 4 м соответственно. Одновременно тот же диктофон синхронно записывал акустический сигнал опорного канала, микрофон которого находился на расстоянии 1 м от акустической колонки. Сигналы микрофонов дискретизировались с частотой 16 кГц.

Обработка заключалась в шумоочистке сигнала основного канала с использованием различных алгоритмов шумоподавления. В качестве опорного брались как сигнал, записанный с микрофона, расположенного около колонки, так и оцифрованные исходные тестовые моносигналы (таким способом моделировалась асинхронная запись сигнала из другого источника). Для количественной характеристики уровня подавления помех использовалась характеристика „уровень подавления шума“ NR (дБ) [3]:

$$NR\ (\text{дБ}) = 1/K \sum_{k=1}^{K} 10\log_{10}(R_k),$$

где $R_k = \sum_{i=1}^{M} x_k^2(i) \Big/ \sum_{i=1}^{M} y_k^2(i)$ — уровень подавления шума на k-м кадре; K — общее количество кадров; M — размер кадра в отсчетах; $x_k(i)$ и $y_k(i)$ — входной сигнал основного ка-

нала и выходной (очищенный от шума) сигнал на k-м кадре соответственно. Усредненные результаты по всем трем видам помех (музыка, речь, шум) приведены на рис. 2.

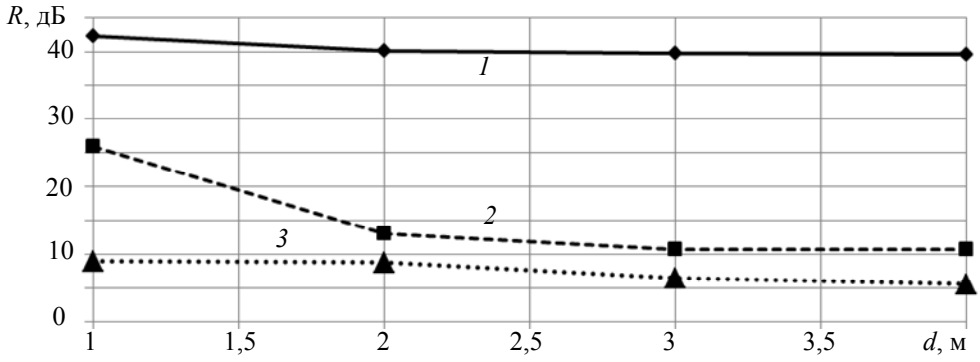


Рис. 2

Отметим, что без процедуры синхронизации как адаптивный линейный компенсатор, так и АСВ дают неудовлетворительные результаты — уровень подавления в обоих случаях практически равен нулю.

Кривая *1* подтверждает, что после синхронизации основного и опорного сигналов АСВ показывает высокую эффективность подавления помехи. С увеличением расстояния между микрофоном и излучателем степень подавления помехи при использовании АСВ снижается незначительно. Кривая *2* иллюстрирует то, что эффективность линейного компенсатора даже в случае синхронной записи помехи оказывается хуже, чем у АСВ, и значительно снижается при удалении микрофона основного канала вследствие уменьшения когерентности помех в опорном и основном каналах. Кривая *3* показывает, что применение адаптивного линейного компенсатора неэффективно в асинхронном случае.

Заключение. Предложен метод шумоподавления для записанных в помещении фонограмм, которые содержат речь, искаженную акустическими помехами, создаваемыми аудиоустройствами. Метод основан на использовании асинхронной аудиозаписи помехи, взятой из стороннего источника — CD, магнитной ленты и т.п. Метод реализуется с использованием действий, требующих участия оператора. Опыт практического применения разработанного метода для шумоочистки реальных фонограмм, поступавших от заказчиков, подтвердил его эффективность.

Центральными моментами метода являются синхронизация сигналов помехи в основном и опорном каналах и алгоритм двухканального спектрального вычитания.

В настоящее время метод встраивается в новую версию редактора Sound Cleaner, продукта ООО „ЦРТ“.

СПИСОК ЛИТЕРАТУРЫ

1. *Aalburg S., Beaugeant C., Stan S., Fingscheidt T., Balan R., Rosca J.* Single-and two-channel noise reduction for robust speech recognition in car // Siemens Corporate Research Report. Siemens AG, ICM Mobile Phones, Multimedia and Video technology, 2002.

2. *Уидроу Б., Стирнз С.* Адаптивная обработка сигналов / Пер. с англ., под ред. *В. В. Шахгильдяна.* М.: Радио и связь, 1981. 440 с.

3. *Bitzer J., Brandt M.* Speech Enhancement by Adaptive Noise Cancellation: Problems, Algorithms and Limits // AES 39th Intern. Conf. Hillerød/Dänemark, 2010. P. 106—113.

4. *Haykin S.* Adaptive Filter Theory. NY: Prentice Hall, 1996. 989 p.

5. *Benesty J., Chen J., Huang Y.* Time Delay Estimation via Linear Interpolation and Cross Correlation // IEEE Transactions on Speech and Audio Processing. 2004. Vol. 12, N 5.

6. *Ignatov P., Stolbov M., Aleinik S.* Semi-Automated Technique for Noisy Recording Enhancement Using an Independent Reference Recording // AES 46th Intern. Conf. Denver, USA, 2012.

7. *Wang L., Nakagava S., Kitaoka N.* Blind Dereverberation Based on Spectral Subtraction by Multi-channel LMS Algorithm for Distant-talking Speech Recognition // IEICE Trans. Inf. Syst. 2011. E94-D(3). P. 659—667.

8. *Бобцов А. А., Колюбин С. А., Пыркин А. А.* Алгоритм управления по выходу с компенсацией синусоидального возмущения для линейного объекта с параметрическими и структурными неопределенностями // Науч.-техн. вестн. информационных технологий, механики и оптики. 2012. № 3 (79). С. 68—72.

9. *Nasu Y., Shinoda K., Furui S.* Cross-channel spectral subtraction for meeting speech recognition // Proc. ICASSP. 2011. P. 4812—4815.

Сведения об авторах

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 621.391.037.372

С. В. Алейник, К. К. Симончик

АЛГОРИТМЫ ВЫДЕЛЕНИЯ ТИПОВЫХ ПОМЕХ И ИСКАЖЕНИЙ В РЕЧЕВЫХ СИГНАЛАХ

Исследованы способы выделения типовых аддитивных помех в системах обработки речевых сигналов. Проведена экспериментальная оценка влияния того или иного детектора помех на эффективность системы верификации диктора. Предложены усовершенствованные алгоритмы выделения помех.

Ключевые слова: *шум, акустические помехи, импульсные помехи, обработка речевых сигналов.*

Введение. Акустические речевые сигналы зачастую искажены аддитивными помехами, значительно снижающими эффективность систем верификации диктора. В общем случае данные аддитивные помехи могут быть разделены на две большие группы: стационарные, присутствующие на всем протяжении сигнала (например, широко известный белый и розовый шум), и нестационарные кратковременные, присутствующие на отдельных участках сигнала.

При наличии помех второй группы входные сигналы редко бывают полностью искажены. Незначительно искаженные участки сигнала чередуются с участками, сильно искаженными импульсными помехами различных типов: клиппированием, кратковременными электрическими наводками, перегрузками и т.п. Именно эти нестационарные помехи и искажения оказывают наибольшее отрицательное влияние. Соответственно используя детекторы, способные на этапе предобработки с высокой вероятностью обнаруживать подобного рода помехи и искажения (с целью их дальнейшего подавления или исключения из анализа), можно существенно улучшить качество систем обработки речи. Основными типовыми помехами и искажениями, рассматриваемыми в настоящей статье, являются щелчки, перегрузки, короткие тональные сигналы, клиппирование.

Следует также отметить, что важными дополнительными требованиями к таким детекторам являются высокая скорость и низкая ресурсоемкость, т.е. типовые требования, предъявляемые к устройствам предобработки.

Щелчки. Несмотря на кажущуюся простоту, обнаружение щелчков представляет собой определенные трудности, поскольку короткие импульсы, воспринимаемые человеком на слух как „щелчки“, могут в общем случае существенно различаться как во временном, так и в частотном представлении (рис. 1, *1* — короткий „классический“ высокочастотный щелчок; *2* — низкочастотный щелчок; *3* — щелчок с короткими осцилляциями; *4* — „длинный“ щелчок с шумовым или осциллирующим заполнением).

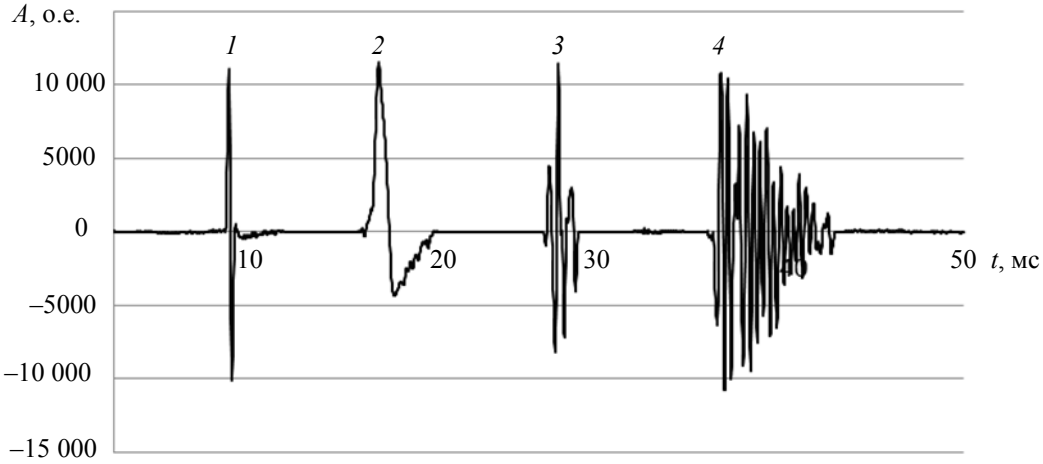


Рис. 1

Например, короткий высокочастотный щелчок хорошо обнаруживается следующим способом. Анализируемый сигнал $x(i)$, где i — дискретный временной индекс, вначале пропускается через высокочастотный (ВЧ) фильтр с частотой среза порядка 2—4 кГц. Затем вычисляется первая разность $d(i) = y(i) - y(i - 1)$, где $y(i)$ — сигнал на выходе фильтра, далее ее абсолютная величина сравнивается с пороговым значением. К сожалению, данный способ не работает на низкочастотных (НЧ) щелчках (кривая *2*), так как, во-первых, основная часть их энергии сосредоточена в низкочастотной области и „срезается“ ВЧ-фильтром, а во-вторых, значение $d(i)$ щелчков данного вида и речевых сигналов различается несущественно.

Результаты исследований различных алгоритмов, основанных на методах линейного предсказания и авторегрессионных моделях [1, 2] показали их высокую вычислительную сложность, поэтому авторы разработали более простой алгоритм обнаружения щелчков различных типов (рис. 2, сплошная кривая — участок анализируемого сигнала со щелчком, пунктир — выходная величина алгоритма (умноженная на 1000 с целью отображения на одном графике с сигналом); t_0—t_3 — временные метки границ окна анализа).

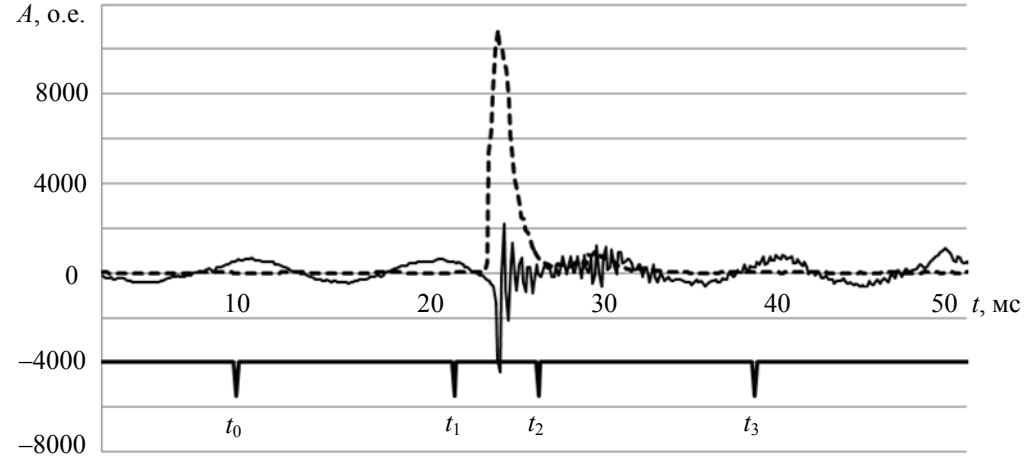


Рис. 2

Разработанный алгоритм включает следующие шаги.

1. Выбирается длина окна анализа (t_0, t_3) таким образом, чтобы выполнялось условие

t_3–$t_0 = KL_c$, где L_c — предполагаемая длительность щелчка и K — масштабный коэффициент, изменяющийся в диапазоне от 10 до 100.

2. Окно разбивается на три части (см. рис. 2), причем длина центральной части выбирается соизмеримой с предполагаемой длиной щелчка, и t_1—$t_0 = t_3$—t_2.

3. Выходная величина V_c, сравниваемая в дальнейшем с пороговым значением, рассчитывается как:

$$V_c(t_{\text{center}}) = \frac{2(t_1 - t_0)}{t_2 - t_1} \frac{\sum_{t=t_1}^{t_2} x^2(t)}{\sum_{t=t_0}^{t_1} x^2(t) + \sum_{t=t_2}^{t_3} x^2(t)}, \quad (1)$$

где $x(t)$ — анализируемый сигнал; $t_{\text{center}} = 0{,}5(t_0+t_3)$ — центр интервала $[t_3, t_0]$.

Нетрудно понять, что V_c в (1) есть отношение мощностей сигнала на различных участках, нормированное таким образом, что в случае стационарного сигнала (например, белого шума) $V_c = 1$. Для речевых сигналов полученные значения V_c колебались от нуля до нескольких единиц. Величина $V_c > 8$ сигнализирует о наличии щелчка (строго говоря, конкретное пороговое значение зависит от выбранной допустимой вероятности ложной тревоги и размеров окна анализа и определяется экспериментально).

Очевидно, что длина интервала t_2–t_1 в идеальном случае должна соответствовать длительности щелчка, подлежащего обнаружению, что в реальных условиях труднодостижимо. В проведенных экспериментах установлено, что если это значение находится в пределах нескольких длин щелчка, то результаты детектора также вполне приемлемы. В противном случае, при значительной априорной неопределенности в длительности предполагаемых щелчков, приходится осуществлять перебор.

Путем моделирования были получены следующие временные параметры детектора: интервал t_2–t_1 5 мс; t_1–t_0 и t_3–t_2 — 60 мс. При таких значениях получены хорошие результаты по детектированию типовых щелчков на реальных речевых сигналах.

Следует заметить, что при обнаружении коротких высокочастотных щелчков бывает полезна предварительная фильтрация ВЧ-фильтром с частотой среза 2—4 кГц.

Перегрузки. Перегрузкой называются короткие (1—2 отсчета) скачки сигнала, импульсы или серии подобных импульсов большой амплитуды, вызванные изменением знака сигнала при так называемом „целочисленном переполнении“. Причины перегрузок кроются в следующем. На практике наиболее широко используемый тип квантования при переводе аудиосигналов в цифровую форму — 16-битовое квантование. При таком типе квантования каждый отсчет сигнала представляет собой целое двухбайтовое число в формате “signed short int” (стандарт ANSI), т.е. амплитуда отсчета изменяется от –32 768 до 32 767. В то же время обработка сигнала может выполняться, например, в форматах “long”, “float” или “double”. При этом если число, получившееся после обработки, выходит за пределы интервала [–32 768, 32 767], то при его простом преобразовании к типу “signed short int” (при записи, например, на диск в WAV-формате) произойдет „переброс знака“, и число, например 32 768, преобразуется в –32 768, число –32 769 — в 32 767 и т.д.

Общие выражения для результата могут быть записаны как:

$$\text{if } (x > 32\,767) \text{ then } y = (x \bmod 32\,767) - 32\,768,$$

$$\text{if } (x < -32\,768) \text{ then } y = -(|x| \bmod 32\,768) + 32\,768,$$

где x — число до преобразования, y — результат преобразования, mod — операция вычисления по модулю.

На слух одиночная перегрузка воспринимается как высокочастотный щелчок, а серия подобных щелчков — как резкий громкий треск, существенно ухудшающий как разборчивость речевого сигнала, так и показатели систем обработки речи.

На рис. 3 приведен типичный пример перегрузки, возникшей при преобразовании величины в формате double (время перегрузки 6,68 мс, значение x = 56 981) в двухбайтовый формат signed short int.

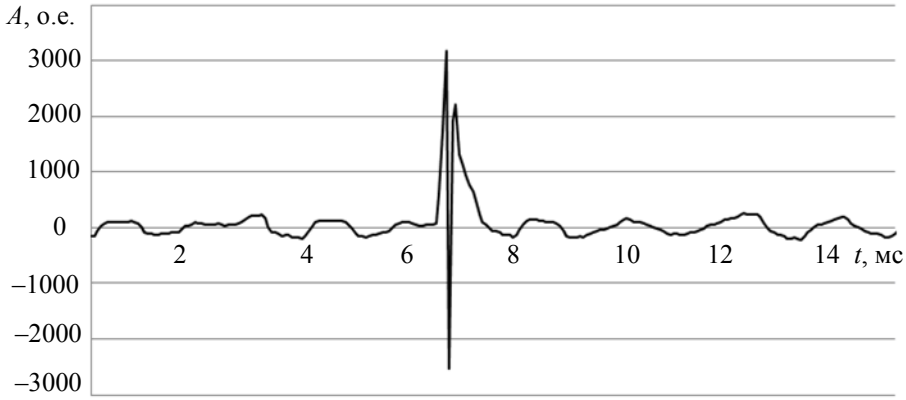


Рис. 3

Одиночная перегрузка (в отличие от серии) с успехом может быть обнаружена с помощью детектора ВЧ-щелчков. Однако, используя первую разность (которая была ранее описана как неэффективная при обнаружении НЧ-щелчков второго типа), возможно создать алгоритм, обнаруживающий как одиночные, так и множественные перегрузки. Дело в том, что „переброс“ знака вызывает сильные резкие скачки амплитуды за один отсчет, часто соизмеримые с динамическим диапазоном сигнала. В данном случае коэффициент вычисляется следующим образом:

$$d(i) = \frac{|x(i) - x(i-1)|}{A_{\max} - A_{\min}}, \qquad (2)$$

где $A_{\max}$ и $A_{\min}$ — максимальное и минимальное значения амплитуды сигнала, вычисленные по всей выборке. Теоретически $0 \le d(i) \le 1$, однако на чистой речи, без перегрузок, величина $d(i)$, как правило, значительно меньше единицы.

Наши эксперименты по определению плотности распределения коэффициента $d(i)$ на большом наборе речевых сигналов показали, что при пороге $T_d = 0,7$ и принятии решения о наличии перегрузки по условию $d(i) > T_d$ вероятность ошибки первого рода (вероятность принять речь за перегрузку) равна приблизительно 10^{-8} на один отсчет сигнала, что дает хорошие результаты даже на длинных сигналах.

Алгоритм детектирования перегрузок представлен ниже.

1. Выбирается величина порога T_d, например, 0,7.

2. По всей выборке сигнала вычисляются его максимальное $A_{\max}$ и минимальное $A_{\min}$ значения.

3. Для каждого отсчета сигнала $x(i)$, $i = 1, N-1$ (здесь N — полная длина сигнала) по формуле (2) вычисляется коэффициент $d(i)$.

Производится сравнение $d(i)$ с выбранным ранее порогом, и в случае $d(i) > T_d$ принимается решение о наличии перегрузки.

Короткие тональные сигналы — это широко известные сигналы телефонного вызова, представляющие собой обычно одну или две гармоники длиной около одной секунды. Отличительной особенностью таких сигналов является высокий уровень и стабильность частоты составляющих гармоник. Соответственно в подавляющем большинстве алгоритмов обнаружения тонов используется анализ спектров мощности (или модулей спектров мощности)

сигналов [3, 4]. Отметим, что тональные сигналы без примеси постороннего шума или в сумме с шумом малой мощности могут быть также с успехом обнаружены детектором клиппированных сигналов, базирующемся на анализе гистограммы [5].

Нами были исследованы два алгоритма обнаружения коротких тонов: на основе подсчета локальных максимумов в спектре и детектор оценки постоянства амплитуды спектральных максимумов. Детектор на основе подсчета локальных максимумов использует тот факт, что при наличии в сигнале тональной компоненты большой амплитуды спектр мощности такого сигнала имеет ярко выраженный узкий пик.

Алгоритм детектирования следующий.

1. Выбирается величина M — длина сегмента сигнала для вычисления спектра мощности.

2. Для каждого сегмента сигнала длиной M вычисляется модуль мгновенного спектра мощности $S(m)$, где $m = 0, M/2$ — дискретная частота.

3. Для всех $m = 0, M/2$ находится спектральный максимум $S_{\max}$.

4. Вычисляется пороговый уровень $T_s = T_{s0} S_{\max}$.

5. Для всех $m = 0, M/2$ подсчитывается целевая величина K_s — количество спектральных отсчетов, превышающих уровень T_s, т.е.: $K_s = \sum_{m=0}^{M/2} k_s$, где

$$k_s = \begin{cases} 1 \;\; if \;\; S(m) \ge T_s, \\ 0 \;\; if \;\; S(m) < T_s. \end{cases}$$

6. Производится сравнение: если $K_s \le 3$, то принимается решение о наличии тональной составляющей в исследуемом фрагменте сигнала.

В алгоритме оценки постоянства амплитуды спектральных максимумов используется тот факт, что на соседних сегментах сигнала амплитуда тональной составляющей изменяется незначительно. В данном алгоритме сравниваются максимумы модулей спектров мощности двух соседних сегментов сигнала $S^j_{\max}$ и $S^{j+1}_{\max}$ (где j — индекс сегмента) и вычисляется их относительная разность:

$$D_s = \frac{\left| S^{j+1}_{\max} - S^j_{\max} \right|}{S^j_{\max}}.$$

Сравнение величины D_s с заранее выбранным порогом T_d дает искомый результат: если $D_s < T_d$, то принимается решение о наличии тональной составляющей в *j*-м фрагменте сигнала.

Пороговые величины T_{s0} и T_d были определены нами в ходе моделирования: $T_{s0} = 0,01$ и $T_d = 0,001$.

Клиппирование — искажение формы сигнала, происходящее при перегрузке усилителя и при выходе выходного напряжения усилителя из его динамического диапазона. На осциллограмме клиппирование обычно выглядит как ограничение сигнала по амплитуде.

На слух клиппирование воспринимается как появление излишней звонкости, „металлического“ звучания и может существенно снижать качество обработки речи.

Алгоритм детектирования клиппирования на основе анализа гистограммы сигнала приведен в работе [5].

Экспериментальная оценка эффективности разработанных алгоритмов. Эффективность предложенных алгоритмов была оценена в ходе экспериментов на примере системы верификации диктора на основе *i*-векторов, описанной в работе [6].

Для тестирования алгоритмов выделения типовых помех использовались записи телефонных разговоров в стандартном GSM-канале: 610 фонограмм различной длительности. Тестовые фонограммы поступали на вход блока предобработки, содержащего параллельно соединенные детекторы: участки фонограмм, на которых срабатывал хотя бы один из включенных детекторов, исключались из дальнейшего анализа. Показателем качества системы был выбран равновероятный уровень ошибок первого и второго рода (Equal Error Rate, EER), широко применяемый для оценки эффективности биометрических систем. Результаты экспериментов представлены в таблице.

Алгоритм детектирования				EER, %
щелчков	перегрузок	клиппирования	тональных помех	
–	–	–	–	13,6
–	–	–	+	10,4
–	–	+	–	10,4
–	+	–	–	10,91
+	–	–	–	10,85

Из таблицы видно, что при отсутствии детекторов (первая строка) качество системы наихудшее (высокий EER). Включение какого-либо детектора приводит к уменьшению EER, т.е. к повышению качества верификации. Следует отметить одинаковое улучшение при работе детекторов клиппирования и тональных помех. По мнению авторов, данный эффект был вызван тем, что, во-первых, в тестовых фонограммах клиппирование практически отсутствовало (в отличие от тональных сигналов телефонных вызовов). И, во-вторых, как уже отмечалось ранее, детектор клиппирования с успехом обнаруживает тональные сигналы, состоящие из одной гармоники.

Заключение. В статье рассмотрены алгоритмы обнаружения типовых помех, наиболее часто встречающихся при обработке речевых сигналов. Указаны характеристики данных алгоритмов, полученные путем моделирования на реальных записях речи. С помощью экспериментального исследования показано, что обнаружение и исключение из анализа речевых сигналов участков с помехами или искажениями способно повысить качество систем верификации диктора.

Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации.

СПИСОК ЛИТЕРАТУРЫ

1. *Esquef P. A. A., Karjalainen M., Välimäki V.* Detection of clicks in audio signals using warped linear prediction // Proc. of the 14th Intern. Conf. on Digital Signal Processing. Greece, 2002. Vol. 2. P. 1085—1088.

2. *Esquef P. A. A., Biscainho L. W. P., Diniz P. S. R., Freeland F. P.* A double-threshold-based approach to impulsive noise detection in audio signals // Proc. EUSIPCO. Finland, 2000. Vol. 4. P. 2041—2044.

3. *So H. C., Chan Y. T., Ma Q., Ching P. C.* Comparison of Various Periodograms for Sinusoid Detection and Frequency Estimation // IEEE Trans. on Aerospace and Electronic Systems. 1999. Vol. 35. P. 945—952.

4. *Grigorakis A.* Application of Detection Theory to the Measurement of the Minimum Detectable Signal for a Sinusoid in Gaussian Noise Displayed on a Lofargram. Research Report, Aeronautical and Maritime Research Laboratory, Melbourne, Australia, 1997.

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

6. *Белых И. Н., Капустин А. В., Козлов А. В., Лоханова А. И., Матвеев Ю. Н., Пеховский Т. С., Симончик К. К., Шулипа А. К.* Система идентификации дикторов по голосу для конкурса NIST SRE 2010 // Информатика и ее применения. 2012. Т. 6, № 1. С. 91—98.

Сведения об авторах

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 656.25-52:656.22.05

С. В. Бибиков, М. Е. Маркисонов, С. А. Панасюк

СОВРЕМЕННАЯ МОБИЛЬНАЯ СИСТЕМА ОПОВЕЩЕНИЯ О ПРИБЛИЖЕНИИ ПОЕЗДОВ

Проанализированы системы оповещения работников путевых бригад о приближении подвижного состава. Проведен сравнительный анализ предложенной авторами мобильной системы оповещения и зарубежных систем.

Ключевые слова: *системы оповещения, приближение поезда, виброакустические колебания.*

При возросшей скорости подвижного состава и сокращающихся межпоездных интервалах существующие системы оповещения о приближении поезда не могут надежно решить задачу безопасности рабочих, занятых текущим ремонтом пути.

Рассмотренные в [1, 2] системы оповещения, в том числе „Сирена“, разработанная ООО „НИИАС“, являются вспомогательными системами обеспечения безопасности, так как сохраняют существующий порядок ограждения места производства работ сигналистами, работают по сигналам сигнализации, централизации, блокировки (СЦБ) и не являются средствами персонального оповещения.

На железных дорогах Японии применяются системы мультимедийной мобильной связи, в которой объединены проводные линии связи и радиосвязь [3]. Местоположение подвижного состава фиксируется с помощью сигнала передатчика, установленного на локомотиве, или на посту СЦБ, поступающего на приемник ретранслятора, который, в свою очередь, посылает сигнал по радиоканалу на портативное устройство. Каждый работник снабжен таким устройством и может оперативно получить всю информацию по обстановке.

Семейство независимых от сигналов СЦБ систем оповещения представляют Minimel 95, разработанная компанией Schweizer Electronic (Швейцария) [4], и система оповещения на основе радиосвязи Autoprowa® фирмы ZÖLLNER — Signal System Technologies [5]. Minimel 95 может быть сконфигурирована как полностью автоматическая, полуавтоматическая, управляемая сигналами или вручную. Модули системы Autoprowa® могут соединяться кабелями либо обмениваться сигналами по радиоканалу. В результате испытаний системы Autoprowa® на Октябрьской железной дороге были выявлены следующие недостатки: система не обеспечивает оповещение операторов дефектоскопных тележек [6]; использование системы малочисленной ремонтной бригадой затруднительно; датчики срабатывают только в момент

проезда поезда мимо них, что приводит к необходимости ограждения на большом расстоянии от участка работ.

По своим свойствам наиболее близок к системам, не зависящим от СЦБ, отечественный комплекс „КОБРА“ (разработка группы компаний „ТВЕМА“ [7]), но вопросы сертификации его надежности и безопасности пока не решены.

В настоящее время специалистами ОАО „НИИАС“ разработана координатная система контроля и оповещения на основе спутниковых радионавигационных систем ГЛОНАСС/GPS [8]. Источником информации о приближении подвижной единицы к месту работ является бортовой оповещатель. В качестве канала передачи данных используется GSM. Преимуществом системы является автономность от сигналов СЦБ. Недостатками — необходимость наличия сети GSM, а также оснащения локомотивного парка и специального самоходного подвижного состава бортовыми оповещателями.

С 2009 г. в ООО „ЦРТ“ ведется разработка переносного сигнализатора для оповещения работников путевых бригад о приближении поезда. Разработано автономное переносное устройство оповещения „Сигнализатор П“, в котором используется принцип анализа виброакустических колебаний, создаваемых в рельсах приближающимся поездом. Устройство не зависит ни от сигналов СЦБ, ни от наличия оповещателя на приближающемся поезде. „Сигнализатор П“ прошел опытную эксплуатацию на Октябрьской железной дороге в 2011 г. Устройство устойчиво обнаруживает приближающийся поезд на бесстыковом пути за 90—120 с до его проезда по месту установки. Но если между поездом и местом установки устройства имеются неоднородности пути (мосты, эстакады либо стрелочные переводы), виброакустический сигнал значительно ослабевает. Интервал времени от момента обнаружения приближающегося поезда до его проезда по месту установки устройства оповещения уменьшается до неприемлемых значений. Это является физическим ограничением принципа работы.

Существует несколько путей решения задачи безопасности оповещения.

1. Обнаруживать приближающийся поезд за неоднородностью пути: повысить качество принимаемого сигнала, чувствительность датчиков, улучшить обработку сигнала до момента аналого-цифрового преобразования.

2. Усовершенствовать алгоритмы обнаружения для принятия решения по слабому сигналу в условиях сильных шумов.

3. Вместо одиночного устройства оповещения использовать распределенную систему датчиков.

Разработанная в ЦРТ оригинальная конструкция узла датчика, снимающего виброакустический сигнал с шейки рельса, значительно повысила чувствительность и соотношение „сигнал—внешний акустический шум“. Переработаны входные цепи устройства, введено раздельное усиление входного сигнала по двум полосам для уменьшения влияния шума. Однако шум, который постоянно присутствует в рельсе, принимается и преобразуется датчиком вместе с сигналом. Согласно экспериментальным данным, это шум со спектральной плотностью вида $1/F$ (розовый шум) с усиленными областями 400—600, 800—1200 Гц и отдельными тонами на более высоких частотах. Его интенсивность различается более чем на 30 дБ в городе и за городом. По рельсам постоянно передаются сигналы СЦБ. Возвратная цепь тягового тока до 400 А также проходит по рельсу. Работающая на путях бригада сама производит различные шумы.

Экспериментальные исследования показали, что сигнал, несущий информацию о приближении поезда и присутствующий во всех ситуациях, порождается шумом качения или взаимодействия колесных пар и рельсов. Вблизи движущегося поезда сигнал шума качения имеет широкий, практически равномерный спектр, вплоть до ультразвуковых частот. Амплитуда сигнала и соотношение спектральных компонент зависят от скорости поезда и качества поверхности рельса. Отличительное свойство шума качения приближающегося поезда — его

специфическое нарастание, зависящее от расстояния, скорости и качества рельсового пути. Зависимость амплитуды сигнала от расстояния до поезда установлена экспериментально:

$$A = K\frac{A_0}{S}, \tag{1}$$

где A — амплитуда сигнала в точке установки датчика, A_0 — исходная амплитуда сигнала в точке нахождения поезда, S — расстояние от поезда до точки установки датчика, K — коэффициент качества пути, определяемый типом шпал и скреплений.

При приближении поезда S можно представить как

$$S = S_0 - Vt,$$

а выражение (1) примет вид:

$$A = K\frac{A_0}{S_0 - Vt}, \tag{2}$$

где S_0 — условное начальное расстояние от датчика до поезда, на котором сигнал становится различимым, V — скорость поезда, t — текущее время.

Коэффициент K является функцией частоты $K = K(f)$, так как рельсовый путь обладает резонансной структурой и различными свойствами для виброакустических колебаний разных частот.

Рассмотрим внимательно выражение (2). Для этого запишем его в виде:

$$A = \frac{K}{V}\frac{A_0}{S_0/V - t}. \tag{3}$$

Параметр S_0/V имеет размерность времени, назовем его *временем захвата сигнала*. Считаем, что первичное обнаружение сигнала, характеристики которого схожи с сигналом приближения поезда, происходит при S_0/V=65—70 с. В этот момент включается алгоритм принятия решения об оповещении. Предполагается, что 15 с достаточно для принятия окончательного решения. Амплитуда сигнала в момент его захвата обратно пропорциональна скорости поезда.

В технической акустике принято правило для шума качения движущегося поезда — его интенсивность нарастает на 9 дБ на каждое удвоение скорости, начиная ориентировочно с 35 км/ч. Это правило не действует для виброакустических колебаний, распространяющихся вдоль рельсового пути. По экспериментальным данным, указанная зависимость интенсивности шума от скорости приближающегося поезда наблюдается только на прямых участках пути. В кривых большого радиуса, на подъемах, спусках и при их комбинациях стабильной зависимости интенсивности шума качения от скорости поезда не наблюдается. Поэтому A_0 в (3) полагаем постоянной величиной.

Изменение характера неравномерности спектра шума при приближении поезда на большом расстоянии от датчика позволило применить соответствующие алгоритмы анализа сигнала. Они дают возможность повысить показатели безопасности оповещения: время и надежность обнаружения подвижного состава. Но в случае физического разрыва рельсового пути на стрелочном переводе виброакустические колебания распространяются только по элементам строения пути, и полезный сигнал практически отсутствует. Таким образом, сфера безопасного применения одиночного устройства оповещения остается достаточно узкой: перегоны бесстыкового пути без неоднородностей либо бесстыковые входные пути станций, также без неоднородностей пути.

С целью расширения сферы применения устройства „Сигнализатор П“ и повышения безопасности авторами предложена система оповещения, которая совмещает в себе преимущества систем с радиоканалом, рассмотренных выше, и датчиков дистанционного обнаружения поезда. Согласно „Эксплуатационно-техническим требованиям на системы оповещения работников, выполняющих работы на перегонах и станциях, о приближении железнодорожного подвижного состава“, применение систем оповещения о приближении подвижного со-

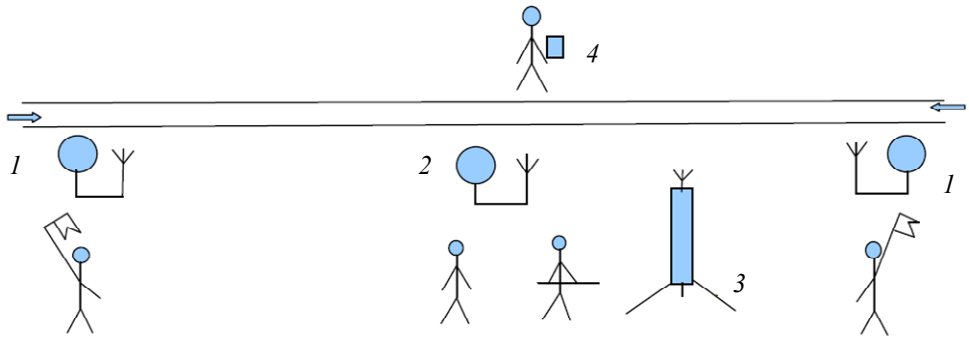
става не отменяет необходимости ограждения мест работы. На рисунке приведена система оповещения, она содержит:

— дистанционные ограждающие датчики приближения поезда с радиоканалом (в количестве до 6 штук);

— контрольный датчик *2* приближения поезда с радиоканалом;

— основной блок обработки и формирования сигнала оповещения *3* с радиоканалом;

— устройства индивидуального оповещения *4* с радиоканалом, работающие в пределах прямой видимости (до 200 м).



Должна быть гарантирована дальность функционирования устройств *1* и *3* не менее 2,5 км.

Датчики *1* и *2* используют принцип анализа виброакустических колебаний, возникающих в рельсе при приближении подвижного состава.

Основной блок *3* содержит устройство обеспечения радиосвязи с датчиками системы, мощную сирену и блок оптической индикации.

Преимущества системы:

— система может продублировать сигналиста в случае возникновения непредвиденных ситуаций;

— промежуточные сигналисты при плохой видимости или на искривленных участках не требуются.

Система может быть усовершенствована так, чтобы постоянное присутствие сигналиста на месте установки удаленного датчика не потребовалось, достаточно только установить и активировать датчик.

Система оповещения (см. рисунок) спроектирована так, чтобы каждый добавляемый элемент повышал надежность системы и безопасность оповещения:

— одиночный основной блок *3* с датчиком *2*, установленным рядом с ним, удовлетворяет всем требованиям к системе оповещения на бесстыковых путях без неоднородностей, поезд обнаруживается гарантированно за 50 с, ограждение места работ не требуется;

— в случае неоднородностей пути или стрелочных переводов в „пятидесятисекундной зоне“ — датчик *1* с радиоканалом, установленный за неоднородностью пути, повышает безопасность оповещения, подавая сигнал раньше, чем это сделает основной блок *3* с установленным рядом с ним датчиком *2*;

— установка промежуточных датчиков *1* повышает надежность системы за счет дублирования;

— оснащение работников путевых бригад, выполняющих задания в отдалении от основного блока *3*, и индивидуальными устройствами *4* повышает безопасность оповещения в условиях протяженного места работ.

Таким образом, создание безопасных систем оповещения, не зависящих от сигналов СЦБ и локомотивных бортовых оповещателей, достигается при использовании распределенных датчиков. В частности, предложенная система оповещения, в которой соединены преимущества систем с радиоканалом и дистанционных датчиков обнаружения приближающегося

поезда, может решить задачу обеспечения безопасности путевых рабочих в широком диапазоне условий применения.

СПИСОК ЛИТЕРАТУРЫ

1. *Щелконогов С. В.* Анализ современных и перспективных систем предупреждения путевых работников о приближении подвижного состава // Молодой ученый. 2012. № 6. С. 61—63.
2. *Ульянов В. М., Меламед Ю. И., Болотин В. И., Жуков В. И., Федосов В. Д.* Автоматическое устройство оповещения о приближении подвижного состава // Автоматика. Связь. Информатика. 2001. № 5. С. 38—42.
3. Система предупреждения при путевых работах // Железные дороги мира. 2003. № 12 [Электронный ресурс]: <http://www.css-rzd.ru/zdm/12-2003/03157.htm>.
4. Minimel 95 [Электронный ресурс]: <http://www.schweizer-electronic.co.uk/products/LOWS-Equipment-MINIMEL95.html>.
5. Autoprowa® — TrackWarningSystems [Электронный ресурс]: <http://www.zoellner.de/index.php/en/produkte/autoprowa>.
6. *Сальникова И.* Россия — не Европа [Электронный ресурс]: <http://zdr.gudok.ru/pub/21/136746/>.
7. Комплекс обеспечения безопасности работ „КОБРА“ [Электронный ресурс]: <http://www.tvema.ru/ru/productList_3781.html>.
8. *Новиков В. Г., Алабушев И. И.* Координатная система контроля и оповещения // Вестн. железнодорожного транспорта. 2008. № 1. С. 45—48.

Сведения об авторах

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

Максим Евгеньевич Маркисонов — ООО „ЦРТ“, Санкт-Петербург; старший менеджер отдела продаж; E-mail: mme@speechpro.com

Сергей Александрович Панасюк — Управление охраны труда, промышленной безопасности и экологического контроля ОАО „Российские железные дороги“, Москва; главный специалист; E-mail: panasyuksa@gmail.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

СИСТЕМЫ СИНТЕЗА РЕЧИ

УДК 81'322.6

А. И. Соломенник, П. Г. Чистиков, С. В. Рыбин,
А. О. Таланов, Н. А. Томашенко

АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ ПОДГОТОВКИ НОВОГО ГОЛОСА ДЛЯ СИСТЕМЫ СИНТЕЗА РУССКОЙ РЕЧИ

Предложены методика и средства автоматизации процедуры создания голоса заданного диктора для работы в системе синтеза речи VitalVoice. Реализованный алгоритм автоматизированной подготовки голоса включает несколько этапов: выбор текстового материала, запись речи с оперативным контролем параметров записи, создание размеченной звуковой базы, настройка параметров подбора элементов.

Ключевые слова: *синтез речи, создание голоса, автоматическая разметка, дифон, корпус текстов.*

Введение. Технология синтеза речи по тексту давно интересует исследователей всего мира. Существуют разные способы получения речевого сигнала: синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный синтез, синтез на основе статистических моделей (HMM-синтез). Синтез методом Unit Selection (выбора элементов, US) [1], подготовка нового голоса для которого составляет предмет настоящей статьи, является одним из видов компилятивного синтеза. Суть его состоит в том, что синтезированная речь компилируется не из базы специально записанных элементов (аллофонов, дифонов, трифонов, полуфонов, слогов и т. п.), каждый из которых представлен единственным вариантом, а из произнесенных предложений естественного языка, и для каждого элемента из множества выбирается наиболее подходящий вариант. Данный метод позволяет достичь очень высокой естественности синтезированной речи. Однако качественный синтез возможен только на основе полного, сбалансированного и корректно размеченного речевого корпуса. С целью разметки речевой базы для метода US в ООО „ЦРТ“ была разработана специальная многоуровневая система [2]. Добавление нового голоса является нетривиальной задачей для любой системы компилятивного синтеза, так как требует записи новой звуковой базы, из которой подбираются элементы, составляющие синтезированную речь. В особенности это актуально для синтеза методом US, поскольку звуковая база для качественного синтеза голоса должна быть достаточно велика (до нескольких часов звучащей речи) [3]. Именно поэтому важно максимально автоматизировать процесс добавления голоса.

В 2010 г. в ООО „ЦРТ“ разработана специальная подсистема [4], на основе которой позднее было создано приложение VoiceConstructor — программа, позволяющая создавать голоса для системы синтеза русской речи VitalVoice [5]. Программа состоит из модулей подготовки текстов, записи фонограмм и формирования звуковой базы голоса.

В **модуле подготовки текстов**, разработанном специально для русского языка, создаются фонетически сбалансированные корпуса текстов заданного размера. Самый простой

способ получить все необходимые для синтеза элементы — записать большую базу данных речи (десятки часов). Но просто наличия большого объема записанной речи недостаточно, корпус должен быть сбалансированным и по возможности полным, т.е. содержать все необходимые единицы во всех возможных контекстах с различными возможными характеристиками, такими как акустические параметры, частота основного тона, длительность, позиция в слоге и т.п. Но так как для создания базы данных нужна сегментация, которая обычно требует по крайней мере некоторой ручной коррекции после автоматической сегментации, размер базы данных влияет на время, необходимое для подготовки ее к использованию. Кроме того, большие базы данных неудобны для хранения и поиска в них. Таким образом, должен соблюдаться баланс между размером и репрезентативностью данных.

Существует целый ряд исследований по автоматическому созданию текстовых корпусов для различных языков [6, 7]. Для русского языка в работе [8] описывается схожий алгоритм. Главное преимущество метода, рассмотренного в работе [9], состоит в том, что он обеспечивает удобство создания текстовой базы, давая возможность не просто выборки предложений из большого корпуса текстов, но позволяет выбрать тип звуковой единицы корпуса, заранее создать и редактировать необходимые корпусы текстов. Модуль автоматической подготовки текстового корпуса был создан на основе программы анализа статистики фонетических единиц [10].

Работа с системой начинается с указания параметров создаваемой базы данных. Пользователь должен выбрать тип основной единицы: дифон или аллофон, установить среднюю скорость речи и желаемый размер базы. Программа показывает текущие размеры базы данных, текста и корпуса. Программа работает с четырьмя корпусами текстов: базовым, включающим в себя наборы частотных и специфических фраз (алфавит, числа, аббревиатуры и т. п.); пользовательским, в который можно загрузить тексты, необходимые для использования в системе синтеза (например, для чтения объявлений в торговом центре имеет смысл ввести примерные тексты объявлений, которые будут подаваться на синтез); фонетическим, который формируется путем выбора предложений из исходного корпуса так, чтобы максимально включить в тексты необходимые для синтеза звуковые единицы (дифоны или трифоны), если их не хватает в базовом и пользовательском корпусах: исходным корпусом, из которого набираются предложения для фонетического корпуса.

Алгоритм генерации фонетического корпуса включает в себя следующие этапы. Во-первых, система транскрибирует все необходимые тексты. Затем вычисляется необходимый объем фонетического корпуса с учетом данных об общем желаемом размере корпуса и размере основного и пользовательского корпусов (если таковые имеются). Предложения выбираются из исходного корпуса в зависимости от количества отсутствующих в создаваемом корпусе единиц, которые они содержат, предложения с максимальным количеством отсутствующих единиц берутся в первую очередь. Если два предложения содержат одинаковое число таких единиц, предложение с менее частотными дифонами будет взято в первую очередь. Также учитывается длина предложения (предпочтение отдается более коротким). Для редких дифонов процедура выбора такая же, она запускается, когда все дифоны исходного корпуса уже присутствуют в основном и пользовательском корпусах. Подбор предложений заканчивается, когда текст достигнет желаемого размера, причем в тот момент, когда в корпус уже добавлены все отсутствующие дифоны, выдается соответствующее предупреждение. Далее на запись подаются предложения из первых трех корпусов.

Модуль записи фонограмм. На этом шаге производится запись звуковых файлов для выбранных текстовых корпусов. Каждое предложение записывается в отдельный файл. Перед проведением сеанса записи требуется измерить шум канала (в паузе). Превышение заданного значения отношения сигнал/шум отмечается индикатором, предупреждающим о том, что следует изменить условия записи, иначе качество создаваемого голоса может оказаться неудов-

летворительным. Аналогичные индикаторы имеются для уровня и для энергии записываемого речевого сигнала. Процесс записи фонограммы контролируется в режиме реального времени с помощью двух графиков: траектории частоты основного тона и осциллограммы сигнала. Траектория частоты основного тона измеряется автокорреляционным методом. Диктор читает предложение за предложением из текущего списка. В любой момент любое предложение можно перезаписать и продолжить запись.

Модуль формирования звуковой базы голоса. На этом шаге производится разметка звуковых файлов, для того чтобы при синтезе из базы голоса выбирались нужные элементы. Метки хранятся в отдельных текстовых файлах, просмотр и корректировка размеченных файлов производятся в звуковом редакторе WaveAssistant. Для формирования базы голоса необходимо получить разбивку на периоды частоты основного тона (ЧОТ) и аллофонную сегментацию.

Для выполнения разметки по ЧОТ в программе WaveAssistant реализован автокорреляционный метод расчета основного тона с предварительной фильтрацией и постобработкой с целью уточнения положения меток основного тона (ОТ). Низкочастотная фильтрация используется для снижения ошибки определения ОТ путем удаления из сигнала составляющих с частотой выше 500 Гц. Высокочастотная предварительная фильтрация используется для определения участков, на которых нет ОТ (невокализованные звуки). Постобработка положения меток позволяет удалять „слишком частые“ или „слишком редкие“ метки, уточнять положение меток в сложных случаях, когда метки смещаются в ту или другую сторону.

Аллофонная сегментация выполняется автоматически с помощью модулей системы распознавания речи (ASR) с использованием HMM (скрытых марковских моделей). Сегментация проводится на основе выравнивания (force alignment) транскрипции и звукового сигнала, она состоит из трех этапов: обучение акустических моделей; сегментация и автоматическая корректировка границ аллофонов. На первом этапе строятся акустические модели монофонов, так как именно монофоны наилучшим образом подходят для данной задачи. Качество сегментации улучшается, если для каждого диктора имеется достаточное количество данных, чтобы обучить индивидуальные модели. Если данных для построения индивидуальных акустических моделей недостаточно, при сегментации используются либо общие акустические модели, построенные по большой базе (более 50 дикторов), либо строятся модели с использованием данных тех дикторов, голоса которых по своим акустическим характеристикам близки к целевому голосу. На шаге сегментации получаются два варианта —„идеальная“ сегментация, которая в точности соответствует заданной транскрипции, и „реальная“ — отличающаяся от первой более точным акустическим соответствием с фонограммой. Оба варианта сегментации в дальнейшем используются при синтезе речи. Заключительный этап автоматической сегментации заключается в автоматической корректировке полученных на предыдущем этапе границ аллофонов на основе дополнительной информации (разметка ЧОТ и правила, составленные на основе статистического анализа систематических неточностей).

Затем выполняется фильтрация звука с целью выравнивания материала по энергии и уменьшения возможной реверберации на глухих участках согласных. Во время сборки базы получаемая для диктора статистическая информация по длительности и амплитуде аллофонов записывается и затем используется при настройке параметров подбора элементов. В зависимости от пола и возраста диктора уточняются настройки различных параметров элементов. Затем пользователю предлагается запустить инсталляцию голоса, по ее завершении новый голос появляется в списке установленных голосов.

Заключение. Рассмотренная методика автоматизированного создания голоса была опробована на речевом материале различного объема (от нескольких минут до 10 часов речи). Она позволила получить практически важные результаты: при минимальной ручной корректировке разметки достигнута почти полная разборчивость речи и практически стопроцентная

узнаваемость исходного диктора даже на базах необходимого объема (от получаса звучащей речи). Реализованный модуль выбора текстового корпуса позволил при том же объеме базы получить большую аллофонную вариативность, что также позволило улучшить получаемую синтезированную речь.

СПИСОК ЛИТЕРАТУРЫ

1. *Black A. W., Hunt A. J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // Proc. of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1. P. 373—376.
2. *Продан А. И., Корольков Е. А., Опарин И. В., Таланов А. О.* Особенности использования многоуровневой разметки звукового корпуса Unit Selection в системе гибридного синтеза „Живой голос“ // Матер. Междунар. конф. „Диалог“. 2009. С. 415—419.
3. *Black A. W.* Perfect Synthesis for all of the people all of the time // Keynote. IEEE TTS Workshop. Santa Monica, CA, 2002. P. 146—170.
4. *Продан А. И., Таланов А. О., Чистиков П. Г.* Система подготовки нового голоса для системы синтеза „Живой голос“ // Матер. Междунар. конф. „Диалог“. 2010. С. 394—399.
5. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th Intern. Conf. on Speech and Computer. SPECOM 2007. Moscow, Russia, 2007. P. 603—608.
6. *Chevelu J., Barbot N., Boeffard O., Delhay A.* Comparing set-covering strategies for optimal corpus design // Proc. of the 6th Intern. Language Resources and Evaluation. 2008. P. 2951—2956.
7. *van Santen J. P. H., Buchsbaum A. L.* Methods for optimal text selection // Proc. of Eurospeech. Rhodes, Greece, 1997. P. 553—556.
8. *Кривнова О. Ф., Захаров Л. М., Строкин Г. С.* Подбор текстового материала и статистический инструментарий для создания речевых корпусов // Сб. тр. XI сессии Российского акустического общества. Т. 3. Акустика речи. Медицинская и биологическая акустика. М.: ГЕОС, 2001. С. 87—92.
9. *Solomennik A. I., Chistikov P. G.* Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system // Матер. Междунар. конф. „Диалог“. 2012. С. 607—615.
10. *Смирнова Н. С., Чистиков П. Г.* Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задач в области речевых технологий // Матер. Междунар. конф. „Диалог“. 2011. С. 632—643.

Сведения об авторах

Анна Ивановна Соломенник — ООО „Речевые технологии“, Минск; научный сотрудник; E-mail: solomennik-a@speechpro.com

Павел Геннадьевич Чистиков — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: chistikov@speechpro.com

Сергей Витальевич Рыбин — канд. физ.-мат. наук; ООО „ЦРТ“, Санкт-Петербург; ведущий программист; Санкт-Петербургский национальный исследовательский университет информационных технологий, кафедра речевых информационных систем; доцент; E-mail: rybin@speechpro.com

Андрей Олегович Таланов — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела синтеза речи; E-mail: andre@speechpro.com

Наталья Александровна Томашенко — ООО „ЦРТ“, Санкт-Петербург; младший научный сотрудник; E-mail: tomashenko-n@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 519.688

П. Г. Чистиков, Е. А. Корольков, А. О. Таланов, А. И. Соломенник

ГИБРИДНАЯ ТЕХНОЛОГИЯ СИНТЕЗА РЕЧИ НА ОСНОВЕ СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ И АЛГОРИТМА UNIT SELECTION

Рассматриваются особенности построения системы синтеза русской речи с использованием двух наиболее распространенных подходов — статистического, на основе скрытых марковских моделей, и конкатенативного, на основе алгоритма Unit Selection. Для решения задачи моделирования интонации разработана методика создания модели голоса русскоязычного диктора. Эксперименты показывают повышение естественности звучания синтезируемой речи.

Ключевые слова: *синтез речи, скрытые марковские модели, Unit Selection, модель голоса.*

Введение. Синтез речи по тексту представляет собой автоматический перевод последовательности символов произвольного текста в соответствующую им последовательность отсчетов звукового сигнала [1—3]. Существует несколько подходов к организации автоматического синтеза речи по тексту. К основным можно отнести синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный, синтез на основании статистических моделей [4—8].

Наиболее распространены в настоящее время подходы, основанные на алгоритме Unit Selection (US) и на скрытых марковских моделях (СММ-синтез). Первый позволяет достичь максимальной естественности звучания синтезированной речи при использовании корректно отсегментированной на разных уровнях сбалансированной речевой базы данных большого объема. В то же время статистический подход, обеспечивая меньшую естественность звучания синтезированной речи (эффект роботизированности), обладает следующими преимуществами:

1) позволяет легко модифицировать характеристики голоса с помощью адаптации/интерполяции моделей дикторов, в то время как алгоритм US позволяет получить речь, стиль которой не отличается от стиля речевой базы;

2) звучание речи, полученной на основе СММ-технологии, естественно, однако в ней отсутствуют резкие, не обусловленные контекстом перепады по частоте и энергии, обычно присущие конкатенативному синтезу. Кроме того, при применении алгоритма US результат синтеза может существенно ухудшиться в случае отсутствия подходящего звукового элемента в базе данных. При использовании моделей отсутствующие в обучающей выборке звуковые элементы синтезируются на основе средних значений, максимально приближенных к требуемым, благодаря применению технологии кластеризации контекстов, основанной на деревьях. Это позволяет добиться разборчивости синтезированной речи в условиях ограниченного количества звуковых единиц в различных контекстах;

3) позволяет разрабатывать новый голос за гораздо меньшее время, а также требует значительно меньше памяти для хранения речевой базы.

В предлагаемой гибридной системе используются оба подхода: оптимальная последовательность звуковых элементов подбирается из речевого корпуса диктора по классическому алгоритму Unit Selection, но с применением статистической интонационной модели, обученной на той же базе, что позволяет повысить естественность звучания синтезируемой речи по сравнению с реализацией на US или только на основе СММ-технологии.

Описание системы. Функционально и структурно систему можно разделить на подсистемы подготовки звуковой базы данных (подготовительный этап) и синтеза речи (рис. 1). Звуковая база данных строится на основе речевого корпуса, состоящего из совокупности звуковых

файлов, каждый из которых содержит запись одного предложения, и соответствующего ему набора файлов разметки, содержащих необходимую информацию о представленных в предложении звуковых единицах [9—12]. По файлам разметки строится индексная база, обеспечивающая быстрый поиск по целевым характеристикам, таким как имя аллофона, имена аллофонов слева и справа, коэффициенты MFCC (Mel-Frequency Cepstral Coefficients) на границах аллофона, энергия на границах, частота основного тона на границах и длительность аллофона.

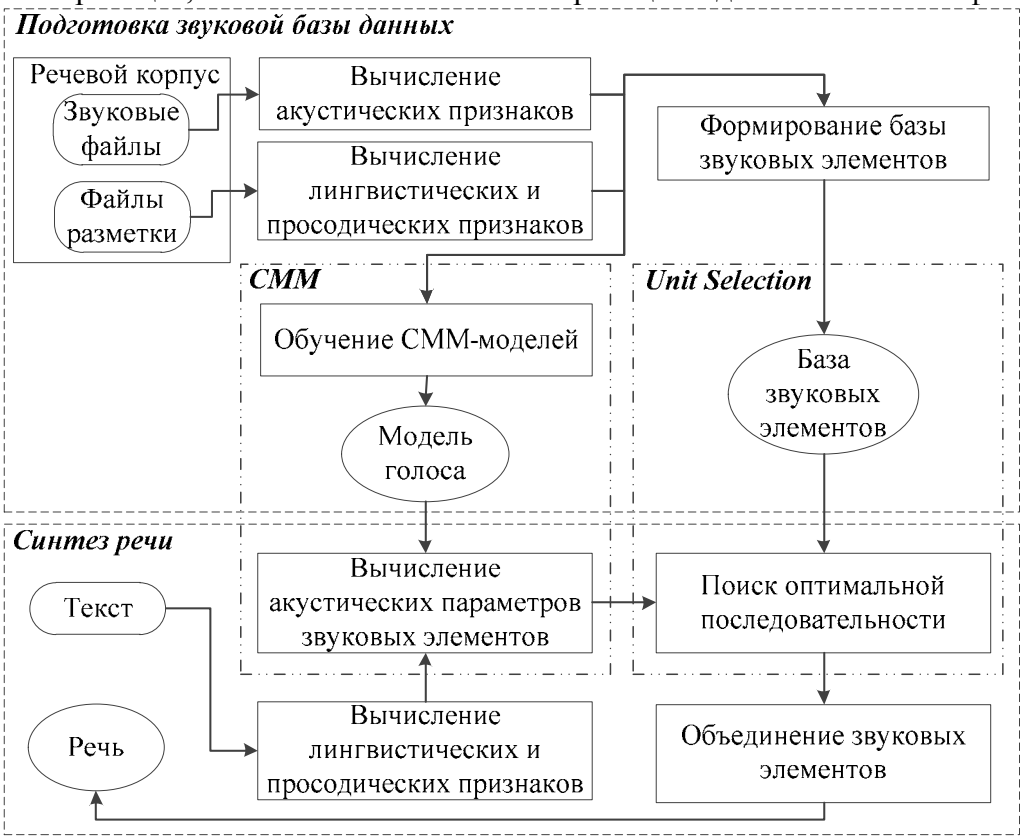


Рис. 1

Процедура моделирования параметров голоса начинается с расчета набора характеристик для всех звуковых файлов [13, 14]. Каждый такой набор описывает короткий участок сигнала (кадр) длительностью 25 мс. В качестве характеристик используются следующие параметры.

— Последовательность $\{c_1, \ldots, c_K\}$ векторов MFCC коэффициентов [15], каждый вектор состоит из 25 коэффициентов и характеризует спектральную огибающую сигнала на фрейме; K — общее количество фреймов.

— Последовательность $\{F0_1, ..., F0_K\}$ значений частоты основного тона.

На следующем шаге для каждого аллофона на основе файлов разметки вычисляется набор лингвистических и просодических признаков, включающий в себя 7 аллофонных (имена стоящего перед предыдущим, предыдущего, текущего, следующего и следующего за следующим аллофонов; позиция от начала и конца слога), 13 слоговых, 8 словных и 3 синтагматических признака.

Далее для каждого аллофона создаются прототипы CMM-моделей. Каждая модель имеет N состояний, допустимы переходы в себя или в следующее состояние. В предлагаемой системе $N = 5$. Каждый выходной вектор наблюдений $\overline{\mathbf{o}}^i$ состоит из четырех потоков $\overline{\mathbf{o}}^i = [\overline{\mathbf{o}}_1^{iT}, \overline{\mathbf{o}}_2^{iT}, \overline{\mathbf{o}}_3^{iT}, \overline{\mathbf{o}}_4^{iT}]^T$: первый содержит значения MFCC, их первых и вторых производных, второй — значение частоты основного тона (ЧОТ), третий — значение первой производной, четвертый — второй производной ЧОТ.

Плотность вектора наблюдений $\bar{\mathbf{o}}^i$ на выходе из состояния n СММ-модели задается следующим выражением:

$$\beta_n(\bar{\mathbf{o}}^i) = \prod_{j=1}^{4}\left[\sum_{l=1}^{R_j}\omega_{njl}\mathcal{N}^{\circ}(\bar{\mathbf{o}}_j^i;\mu_{njl},\Sigma_{njl})\right],$$

где $\mathcal{N}^{\circ}(\cdot;\mu,\Sigma)$ — плотность нормального распределения с вектором средних значений μ и матрицей ковариации Σ, ω_{njl} — весовой коэффициент l-й компоненты смеси j-го потока выходного вектора n-го состояния, R_j — количество компонент в смеси j-го потока. Для k-й СММ-модели вектор длительности состояний $\bar{d}_k = [\bar{d}_{k1},...,\bar{d}_{kN}]^T$ моделируется N-мерным однокомпонентным гауссовым распределением. Вероятности выходных значений моделей спектральных параметров (MFCC и ЧОТ) и длительностей переоцениваются при помощи алгоритма Баума—Велша [16].

Процесс построения модели голоса завершается кластеризацией состояний СММ-моделей на основе деревьев решений. На данном шаге генерируются параметры отсутствующих в обучающей речевой базе элементов, что, в свою очередь, обеспечивает синтез разборчивой речи даже при небольшом объеме обучающего материала. Итоговая интонационная модель голоса состоит из N+1 деревьев: N — для хранения по каждому из состояний параметров СММ-модели ЧОТ вместе с первой и второй производными, и одно — для параметров СММ-модели длительностей.

На вход системе синтеза подается текст без какой-либо предварительной ручной обработки. На основе текстовой информации для каждого предложения формируется целевая последовательность аллофонов и вычисляются лингвистические и просодические признаки для каждого из них. Тип и структура признаков аналогичны тем, что используются на этапе подготовки звуковой базы данных. На основе этой информации по модели голоса определяются акустические признаки каждого аллофона: значения ЧОТ, энергии и длительности. По рассчитанным акустическим и лингвистическим характеристикам из речевой базы выбирается группа наиболее подходящих звуковых элементов. Для того чтобы определить, насколько тот или иной элемент базы подходит для синтеза данной единицы, вводятся понятия стоимости замены (target cost) и стоимости связи (concatenation cost) [17].

Стоимость замены для элемента из базы u_i по отношению к искомому элементу t_i вычисляется по формуле:

$$C_t(u_i,t_i) = \sum_{k=1}^{p} w_{tk} C_{tk}(u_i,t_i),$$

где C_{tk} — расстояние между k-ми характеристиками элементов, w_{tk} — вес для k-й характеристики. Другими словами, стоимость замены есть взвешенная сумма различий в признаках между целевым (требуемым) элементом и конкретным элементом речевой базы. В качестве признаков могут выступать ЧОТ, длительность, контекст, позиция элемента в слоге, слове, количество ударных слогов во фразе и др.

Выбранные элементы должны не только мало отличаться от целевых, но и хорошо соединяться друг с другом. Стоимость связи двух элементов может быть определена как взвешенная сумма различий в признаках между двумя последовательно выбранными элементами:

$$C_c(u_{i-1},u_i) = \sum_{k=1}^{q} w_{ck} C_{ck}(u_{i-1},u_i),$$

где C_{ck} — расстояние между k-ми характеристиками элементов, w_{ck} — вес для k-й характеристики.

Общая стоимость целой последовательности из n элементов есть сумма введенных выше стоимостей:

$$C(u,t)=\sum_{i=1}^{n}C_t(u_i,t_i)+\sum_{i=2}^{n}C_c(u_{i-1},u_i). \qquad (1)$$

Задача алгоритма Unit Selection — выбрать такое множество, которое бы минимизировало общую стоимость согласно формуле (1).

В завершение происходит объединение выбранной последовательности элементов в единый звуковой поток, на выходе представляющий собой синтезированную речь.

Экспериментальные результаты. Примеры работы системы представлены на рис. 2—4, на которых приведены соответственно осциллограммы, спектрограммы и графики динамики частоты основного тона для фразы „Это очень важно!“. На приведенных рисунках в верхней части представлены данные для фразы, записанной реальным диктором, а в нижней — для ее синтезированного варианта. Следует отметить, что синтезируемая фраза не была включена в обучающую выборку.

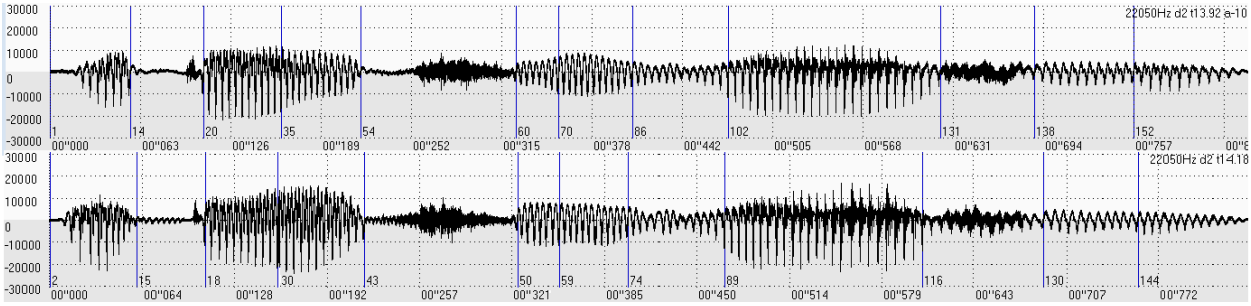


Рис. 2

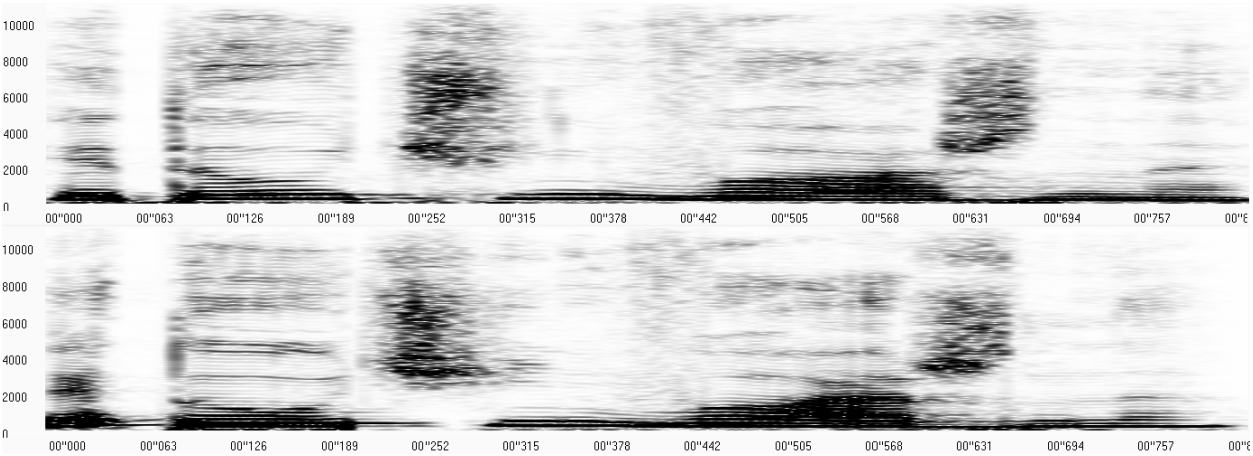


Рис. 3

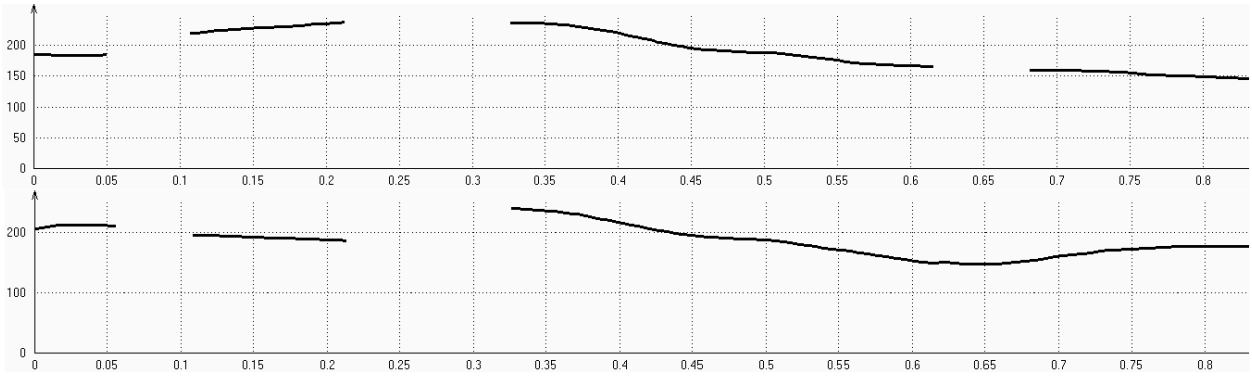


Рис. 4

На основе данных диаграмм можно сделать вывод, что синтезированная фраза имеет практически такие же темпоритмические и спектральные характеристики, как и ее эквивалент, произнесенный диктором, что достигается за счет определения значений этих характеристик на основе скрытых марковских моделей.

Ниже приведены результаты сравнения показателей естественности речи (значения в интервале от 0 до 5, где 5 — максимальная оценка естественности) представленной в работе системы с системой на основе метода US, лежащей в основе гибридного подхода. Пять экспериментов оценивали два (мужской и женский) голоса, данные в таблице усреднены. Как видно из результатов эксперимента, применение гибридного подхода позволило улучшить показатели естественности синтезированной речи.

Тип подхода к синтезу		
Unit Selection	гибридный подход	естественная речь
4,0	4,3	4,8

Заключение. В ходе проведенных исследований была разработана гибридная система синтеза русской речи по тексту, в основе которой лежат скрытые марковские модели и алгоритм Unit Selection. Результаты испытаний показали, что по показателям естественности звучания данная система является одной из лучших среди систем синтеза на русском языке.

СПИСОК ЛИТЕРАТУРЫ

1. *Dines J.* Model based trainable speech synthesis and its applications. Ph. D. Thesis. Brisbame, Australia: Queensland University of Technology, 2003.
2. *Dutoit Th.* Introduction au traitement de la parole // Faculte Polytechnique de Mons. 2002.
3. *Stilianou Y.* Harmonic plus noise models for speech, combined with statisticals methods, for speech and speaker modification. Ph. D. Thesis. Paris, France: Ecole Nationale Superieure des Telecommunications, 1996.
4. *Klatt D. H.* Review of text-to-speech conversion for English // J. of the Acoustical Society of America. 1987. Vol. 82. P. 737—793.
5. *Tokuda K.* HMM-based Speech Synthesis System (HTS). 2011 [Электронный ресурс]: <http://hts.sp.nitech.ac.jp>.
6. *Huang X., Acero A., Adcock J., Goldsmith J., Liu J. W.* A Trainable Text-to-Speech System // Proc. of the Intern. Conf. on Spoken Language Processing. Philadelphia, PA. 1996. Vol. 4. P. 2387—2390.
7. *Donovan R. E., Eide E. M.* The IBM Trainable Speech Synthsis System // Proc. ICSLP'98. Sydney, Australia, 1998.
8. *Donovan R. E., Ittycheriah A., Franz M., Ramabhadran B., Eide E., Viswanathan M., Bakis R., Hamza W.* Current Status of the IBM Trainable Speech Synthesis System // Proc. 4th ESCA Tutorial and ResearchWorkshop on Speech Synthesis. Scotland, UK. 2001.
9. *Продан А., Чистиков П., Таланов А.* Система подготовки нового голоса для системы синтеза "VITALVOICE" // Компьютерная лингвистика и интеллектуальные технологии. 2010. № 9 (16). С. 394—399.
10. *Смирнова Н., Чистиков П.* Программа анализа фонетических статистик в текстах на русском языке и ее использование для решения прикладных задачах в области речевых технологий // Там же. 2011. № 10 (17). С. 632—643.
11. *Чистиков П., Хомицевич О.* Автоматическое определение границ предложений в потоковом режиме в системе распознавания русской речи // Вестн. МГТУ им. Н. Э. Баумана. 2011. Вып. S. С. 117—125.
12. *Chistikov P., Khomitsevich O.* On-line automatic sentence boundary detection in a Russian ASR system // SPECOM 2011 Intern. Conf. 2011. P. 112—117.
13. *Чистиков П.Г.* Моделирование параметров русской речи в системе синтеза // Сб. тез. докл. конгресса молодых ученых. Вып. 2. СПб: НИУ ИТМО. 2012. С. 227—228.
14. *Chistikov P., Korolkov E.* Data-driven Speech Parameter Generation for Russian Text-to-Speech System // Компьютерная лингвистика и интеллектуальные технологии. 2012. № 11 (18). С. 103—111.

15. *Fukada T., Tokuda K., Kobayashi T., Imai S.* An adaptive algorithm for mel-cepstral analysis of speech // Proc. of the IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 1992. P. 137—140.

16. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* Hidden semi-Markov model based speech synthesis // Proc. of the Intern. Conf. on Spoken Language Processing (ICSLP). 2004. P. 1393—1396.

17. *Black A.W., Hunt A.J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // Proc. of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1. P. 373—376.

Сведения об авторах

Павел Геннадьевич Чистиков — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: chistikov@speechpro.com

Евгений Александрович Корольков — ООО „ЦРТ“, Санкт-Петербург; научный сотрудник; E-mail: korolkov@speechpro.com

Андрей Олегович Таланов — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела синтеза речи; E-mail: andre@speechpro.com

Анна Ивановна Соломенник — ООО „Речевые технологии“, Минск; научный сотрудник; E-mail: solomennik-a@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 81'322.6

А. И. Соломенник, А. О. Таланов, М. В. Соломенник, О. Г. Хомицевич, П. Г. Чистиков

ОЦЕНКА КАЧЕСТВА СИНТЕЗИРОВАННОЙ РЕЧИ: ПРОБЛЕМЫ И РЕШЕНИЯ

Рассмотрены различные аспекты проблемы оценки результатов работы систем синтеза речи. Приведен краткий обзор существующих методик оценки качества.

Ключевые слова: *синтез речи, качество синтезированной речи, сравнение систем синтеза речи.*

Введение. Синтезированная речь в последние годы все больше используется в различных сферах, например, в банковских системах голосового самообслуживания, транспортных компаний, при проведении телефонных опросов. Синтезированными голосами „говорят“ мобильные устройства, озвучиваются аудиокниги. Поэтому задача оценки качества синтезированной речи и сравнения систем синтеза между собой становится как никогда актуальной. Однако в этой области существует немало проблем. Основной можно назвать субъективность оценок: кто-то обращает внимание на тембр синтезированного голоса или на ошибки в произношении, для кого-то голос слишком „роботизирован“ или, наоборот, „излишне живой“ и непредсказуемый. В настоящей статье будут рассмотрены существующие подходы к объективной оценке качества синтеза в целом и его отдельных компонентов.

Оценка качества лингвистической обработки. Системы синтеза могут сравниваться и оцениваться по следующим объективным параметрам, отражающим решение задач лингвистической обработки в системах синтеза:

1) выделение предложений в тексте и разбиение их на отдельные слова; разметка текста на буквы, специальные символы, цифры и знаки пунктуации;

2) нормализация текста — расшифровка сокращений, аббревиатур, цифровых обозначений, номеров телефонов, дат, времени и т.п.;

3) определение места ударения и морфо-грамматических характеристик слов в предложении, для этого обычно используется словарь и/или набор правил или статистические модели;

4) снятие омонимии (омографии), т.е. выбор одной из нескольких словоформ, соответствующих тому или иному слову текста. Эти словоформы могут различаться ударением, наличием буквы „ё“ или грамматическими характеристиками;

5) построение сегментной транскрипции по правилам или по словарю (в зависимости от языка).

Для объективной оценки качества лингвистической обработки может быть создан тестовый текст либо использован фрагмент готового текстового корпуса, на основе которого вычисляется процент ошибок по каждому из параметров [1]. Эти оценки могут быть получены автоматически, если для сравнения доступен нормализованный и размеченный системой синтеза текст.

Оценка просодической обработки. К такой обработке относится использование тех компонентов, которые придают тексту интонационное оформление, т.е. происходит деление текста на просодические единицы — синтагмы, определение длины пауз между синтагмами и выбор интонационного оформления для каждой из синтагм. Затем происходит вычисление физических параметров — длительности, частоты основного тона (ЧОТ), энергии — на основе полученных данных. Деление на синтагмы и выбор интонации могут осуществляться как по правилам, так и на основе статистических моделей, причем этап выбора интонационного типа в последнем случае может быть пропущен, система сразу на основе имеющихся данных может переходить к предсказанию требуемых физических параметров звуков.

На этом этапе объективная количественная оценка представляет собой уже менее тривиальную задачу, поскольку вышеуказанные параметры в естественной речи могут варьироваться. При оценке расстановки пауз могут отдельно учитываться места, где пауза необходима, возможна и недопустима [2]. Выбор интонационной модели внутри одной системы обозначений может быть оценен таким же образом (по набору допустимых вариантов), но при сравнении разных систем интонация обычно оценивается уже в выходных звуковых файлах по параметру схожести с естественной речью.

Оценка акустического модуля. Возможные проблемы в работе акустического модуля существенно зависят от технологии синтеза: например, в формантном, аллофонном или дифонном компилятивном синтезе это может быть общая заметная неестественность (роботизированность) звучания одновременно с неудачными отдельными звуками; в компилятивном синтезе методом Unit Selection (US) — различимые стыки звуков, призвуки, несоответствие интонационного оформления логически обусловленному контекстом, причем ошибки обычно неравномерно распределены по тексту; в синтезе, основанном на статистическом моделировании (HMM), — роботизированность всей речи или звуков определенного типа, в то время как резких „скачков“ тона или энергии, как в синтезе US, обычно не наблюдается. Если используется значительная модификация звука, в синтезированной речи появляются заметные призвуки и эффект роботизированности.

Степень влияния результата работы акустического модуля на общее качество синтеза сложно переоценить. Поэтому в технологии синтеза основное внимание уделяется именно развитию технологии получения результирующего речевого потока. Опосредованно оценить качество работы этого модуля можно на основе оценки качества синтеза (или оценки общего впечатления), поскольку он формирует выходной сигнал на основании работы предыдущих модулей.

Оценка общего качества синтеза. Методы оценки качества синтеза можно в первую очередь разделить на две большие группы: субъективные (MOS-оценка) и инструментальные.

К первой относятся разного рода тесты, опросники, заполняемые экспертами — специалистами либо наивными слушателями. При создании опросников обычно используются рекомендации P.85 ITU-T „Метод субъективной оценки качества речи устройств речевого вывода“ [3]. В них используется MOS-оценка по пятибалльной шкале по нескольким категориям: общее впечатление, слуховое усилие, естественность, понимание смысла сообщения, темп, разборчивость, приятность голоса. На основе этих критериев принимается решение о приемлемости голоса (для определенных задач) по двубалльной шкале.

Однако проведение такого тестирования является довольно трудоемкой задачей. Для того чтобы ускорить процесс оценки и сделать его более доступным на каждом шаге разработки систем синтеза, создают различные инструментальные (или объективные) методы оценки качества синтеза. Такие методы основываются как на автоматическом сравнении синтезированной речи (с использованием различных мер близости) с „живой“ речью того же диктора [4, 5], так и на построении дикторонезависимых моделей естественной речи и различных методах оценки того, насколько синтезированная речь к ним приближена [6]. При этом исходным является предположение о том, что в естественной речи невозможны резкие скачки в частоте основного тона, энергии или спектральных составляющих, характерные для систем конкатенативного синтеза. В работе [7] предлагается инструмент, позволяющий оценивать качество просодической обработки на основании данных о значениях ЧОТ и длительности звуков речи.

Кроме упомянутых ранее характеристик речи в системе оценки качества синтезированной речи могут быть использованы следующие признаки, оценка которых может выполняться автоматически [8]:

— интонированность/монотонность определяется по изменению производной ЧОТ;

— ритмичность — параметр, который может характеризовать разные аспекты речи, прежде всего он определяется паузами, разбивающими речь на относительно равномерные отрезки;

— мелодичность — параметр, отражающий долю голосовых (вокализованных) фрагментов речи.

Подходы и системы оценки. В России для объективной оценки качества синтезированной речи чаще всего используется ГОСТ Р 50840-95 „Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости“. Наряду с ним используются различные тесты отдельных компонентов [9], но единого стандарта оценки пока нет. В работе [10] был предложен подход к комплексной оценке систем синтеза русской речи, однако он не имеет широкой известности и практически не применяется.

Одним из главных событий в сфере синтеза речи является Blizzard Challenge — „соревнования“ синтезаторов. Голоса для сравнения систем синтеза создаются на основе одних и тех же звуковых баз данных, предоставляемых перед началом соревнований. По прошествии времени, отведенного на создание голосов, участникам выдается набор текстов, синтезированные звуковые файлы для которых необходимо предоставить организаторам для оценки. В 2010 г. соревнования проводились для корпусов речи на английском и китайском языках [11]. В качестве дополнительного задания в 2012 г. предлагалось разработать собственный метод оценки качества синтеза и провести оценку [12].

По образцу Blizzard Challenge для испаноязычных синтезаторов был организован конкурс Albayzin [13]. Существует стандартизованный набор тестов для синтеза речи на французском языке, разработанный в ходе национального проекта EvaSy (Evaluation of speech synthesis systems — оценка систем синтеза речи) [14].

Заключение. Оценка качества синтезаторов в последние годы является предметом широких исследований; за рубежом активно ведутся работы по стандартизации оценок. Для русскоязычных синтезаторов существуют отдельные перспективные разработки, на основании

которых должен быть выработан единый стандарт качества синтезированной речи. Представленный обзор наработок в этой области за последние несколько лет является первым шагом к выработке такого стандарта.

Система оценки качества синтезированной речи может применяться для решения следующих задач.

Тестирование системы синтеза в процессе разработки. К системе оценки предъявляются следующие требования: она должна быть автоматической; иметь достаточно высокое быстродействие; может оцениваться как на соответствие голосу конкретного диктора, так и на соответствие общим параметрам речевого сигнала. Для анализа должны быть доступны результаты всех этапов синтеза, и проверка должна осуществляться с использованием промежуточной информации, генерируемой системой в явном виде.

Оценка собственной системы синтеза речи в сравнении с конкурентами. Для этого может применяться как автоматическая дикторонезависимая оценка [6], так и оценка экспертов. В данном случае может быть затруднен доступ к результатам синтеза: для коммерческих приложений обычно доступны только интерактивные демоверсии, при помощи которых можно получить образцы звука низкого качества с фоновой музыкой и др. в целях защиты от коммерческого использования, или же доступны только заранее подготовленные примеры. Для корректного сравнения результатов работы синтезаторов необходимо использовать их полнофункциональные версии.

Участие в конкурсах, проводимых независимыми компаниями. Система оценки может быть не автоматической, но автоматизированной. Для оценки системы могут привлекаться большие человеческие ресурсы (например, заинтересованные пользователи Интернета). Хоть внутренняя структура систем синтеза и останется закрытой, будет возможно получение промежуточных результатов работы системы в унифицированном виде. Системы синтеза могут тестироваться на одной и той же голосовой базе, на основе которой строится синтезированный голос.

СПИСОК ЛИТЕРАТУРЫ

1. *Sproat R., Black A. W., Chen S., Kumar S., Ostendorf M., Richards C.* Normalization of non-standard words // Computer Speech and Language. 2001. Vol. 15. P. 287—333.

2. *Хомицевич О. Г., Соломенник М. В.* Автоматическая расстановка пауз в системе синтеза русской речи по тексту // Матер. Междунар. конф. „Диалог“. 2010.

3. Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, ITU-T Rec. Int. Telecom. Union. 1994. 85 p.

4. *Vepa J., King S., Taylor P.* Objective distance measures for spectral discontinuities in concatenative speech synthesis // Proc. Intern. Conf. Spoken Language Processing. September, 2002. P. 2605—2608.

5. *Stylianou Y., Syrdal A.* Perceptual and objective detection of discontinuities in concatenative speech synthesis // Proc. Intern. Conf. Acoustics, Speech, and Signal Processing. June, 2001. P. 837—840.

6. *Falk T. H., Möller S.* Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems // IEEE Signal Proc. Letters. 2008. Vol. 15. P. 781—784.

7. *Norrenbrock C. R., Hinterleitner F., Heute U., Moller S.* Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals // IEEE Signal Proc. Letters. 2012. P. 255—258.

8. *Киселев В. В., Давыдов А. Г., Ткаченя А. В.* Система определения эмоционального состояния диктора по голосу // Междунар. науч.-техн. конф. „Открытые семантические технологии проектирования интеллектуальных систем“ (OSTIS-2012) / Под ред. *В. В. Голенкова*. Минск: БГУИР, 2012. С. 355—358.

9. *Гецэвіч Ю. С.* Алгарытмы лінгвістычнай апрацоўкі тэкстаў для сінтэзу маўлення на беларускай і рускай мовах: Дыс. … канд. тэхн. навук. Мінск, 2012. 191 с.

10. *Русанова О. А.* Исследование и разработка методов анализа и оценки качества синтезированной устной речи. Дис. … канд. техн. наук. Красноярск, 2004. 107 с.

11. [Электронный ресурс]: <http://festvox.org/blizzard/blizzard2010.html>.

12. [Электронный ресурс]: <http://www.synsig.org/index.php/Blizzard_Challenge_2012_Rules>.

13. *Méndez F. et al.* The Albayzín 2010 Text-to-Speech Evaluation // Fala2010. 2010. P. 317—340.

14. [Электронный ресурс]: <http://www.technolangue.net/article.php3?id_article=202>.

Сведения об авторах

Анна Ивановна Соломенник — ООО „Речевые технологии“, Минск; научный сотрудник; E-mail: solomennik-a@speechpro.com

Андрей Олегович Таланов — канд. техн. наук; ООО „ЦРТ“, Санкт-Петербург; руководитель отдела синтеза речи; E-mail: andre@speechpro.com

Михаил Васильевич Соломенник — канд. техн. наук; ООО „Речевые технологии“, Минск; ведущий инженер-программист; E-mail: solomennik-m@speechpro.com

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

Павел Геннадьевич Чистиков — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: chistikov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 519.688

О. Г. Хомицевич, С. В. Рыбин, И. М. Аничкин

ИСПОЛЬЗОВАНИЕ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА ДЛЯ НОРМАЛИЗАЦИИ ТЕКСТА И СНЯТИЯ ОМОНИМИИ В СИСТЕМЕ СИНТЕЗА РУССКОЙ РЕЧИ

Исследована проблема разрешения неоднозначности прочтения различных элементов при работе системы синтеза русской речи по тексту VitalVoice. Описываются особенности использования морфологического и синтаксического анализа при расшифровке сокращений и специальных знаков, а также снятии омонимии (омографии). Данные экспериментов свидетельствуют о том, что выбранные методы позволяют правильно прочесть более 95 % сложных элементов естественного текста.

Ключевые слова: *синтез речи по тексту, синтаксический анализ, морфологический анализ, омонимия, омография, нормализация текста.*

Введение. Система автоматического синтеза речи преобразует текст, подающийся на ее вход, в звучащую речь. Необходимым этапом преобразования текста в речь является подготовка текста к тому, чтобы на его основании могла быть составлена фонетическая транскрипция, использующаяся далее для подбора необходимых звуковых элементов. Так, записи, которые не могут быть прочитаны непосредственно в том виде, в котором они встречаются в тексте (сокращения, цифры, небуквенные значки, элементы других алфавитов, например, латиницы в русском тексте), должны быть в итоге приведены к виду „полноценных“ слов русского языка. Например, в предложении *Порядка 22 % от объема полученных в 2011 г. средств — 172 тыс. долларов — были переданы 39 благотворительным организациям* элементы *22, %, 2011 г., 172, тыс, 39* не могут быть транскрибированы в исходном виде и долж-

ны быть преобразованы в слова *двадцати двух, процентов, две тысячи одиннадцатом году, сто семьдесят две, тысячи, тридцати девяти*. Кроме того, для русского языка при построении транскрипции необходима информация о месте ударения в слове, которое не обозначается на письме и должно быть определено отдельно для каждого слова в предложении.

Основной проблемой, возникающей при решении данных задач, является неоднозначность в прочтении элементов. Например, при расшифровке цифр возможен выбор количественного либо порядкового числительного; так, в вышеприведенном примере *22, 172, 39* — количественные числительные, *2011* — порядковое. Существенной проблемой для русского языка, обладающего богатой морфологией, является выбор правильной формы (падежа, рода, числа) при расшифровке элемента: *двадцать два* ***процента***, но *двадцать один* ***процент***, *двадцать* ***процентов***; ***тридцати девяти*** *благотворительным организациям*, но ***тридцатью девятью*** *благотворительными организациями*. При определении места ударения в слове источником неоднозначности является совпадение написания различных слов (омонимия, в более узком смысле — омография, т.е. совпадение написания слов, различающихся по звучанию).

Для решения проблемы выбора варианта расшифровки элемента, варианта ударения в слове и т.п. применяются различные методы анализа текста. Достаточно популярны статистические методы [1, 2], которые основаны на выявлении закономерностей в тексте путем обучения математических моделей. Недостатком таких методов является необходимость опираться на большие объемы соответствующим образом подготовленных текстов: для обучения автоматической программы расшифровки сокращений, цифр и т.п. требуется большой корпус текстов, где все такие элементы соотносятся с правильной расшифровкой, а для снятия омонимии (омографии) — корпус текстов, включающий разнообразные омографы с указанным правильным прочтением. Для русского языка получение такой текстовой базы является проблематичным.

Возможно также использовать в системе синтеза речи синтаксический и семантический анализ (парсинг) текста [3—5]. Однако полноценный разбор зачастую требует существенных вычислительных ресурсов, что нежелательно для коммерческих систем автоматического синтеза речи, которые должны работать в режиме реального времени или с опережением; к тому же именно неоднозначность многих словоформ языка вызывает наибольшие затруднения для многих синтаксических анализаторов [6].

Для расшифровки специальных обозначений и снятия омонимии при синтезе речи в системе VitalVoice используется частичный (локальный) лингвистический (морфологический и синтаксический) анализ текста, т.е. в процессе работы программы анализируется окружение конкретного слова (цифры, знака…). Дополнительное достоинство данного метода заключается в том, что алгоритм может быть сформулирован в виде контекстных правил с интуитивно понятным синтаксисом, которые содержатся в отдельных файлах, а не в программном коде, и могут оперативно редактироваться лингвистом для настройки работы системы. Экспериментальная проверка показывает, что этот метод позволяет корректно разрешить подавляющее большинство случаев неоднозначности чтения в естественном тексте на русском языке.

Расшифровка сокращений и специальных знаков. В текстах на русском языке, таких как газетные статьи, новостные сообщения, научно-популярная и художественная литература и др., встречаются различные типы специальных обозначений. В первую очередь, это сокращения и условные обозначения из различных элементов (буквы, цифры, небуквенные символы): *км, и.о., мск, Гб, м/с, м^2, C#,* а также специальные знаки: *%, °, $, №*. Помимо необходимости расшифровки трудности создает тот факт, что многие сокращения пишутся с точкой, а значит, их наличие должно быть дополнительно учтено в алгоритме деления текста на предложения.

Расшифровка сокращений и специальных знаков производится за счет анализа соседних, а также других слов предложения. Прежде всего нужно учесть семантическую неоднозначность: многие сокращения имеют разную расшифровку в зависимости от контекста, например, *м.* может обозначать „метр“ или „метро“; *ст.* — „станция“ или „статья“, или совпадать с несокращенными словами, например, *Кб* — „килобайт“ или аббревиатурой КБ („конструкторское бюро“), *им.* — „имени“ или личное местоимение. Для снятия подобной неоднозначности осуществляется поиск слова или другого элемента, ключевого для расшифровки: *2012 г.* („год“), *г. Псков* („город“), *ст. 105 УК РФ* („статья“), *ст. Москва-Сортировочная* („станция“) и т.п. Выбор правильной формы осуществляется при помощи анализа ближайшего контекста слова; основную роль играет наличие числительного слева (*1 км* „километр“*, 2 км* „километра“*, 12 км* „километров“*, 22 км* „километра“) и наличие предлога слева, в том числе перед числительным (*более 1 км* „километра“, *к 1 км* „километру“, *до ст. Бологое* „станции“).

Расшифровка цифровых записей включает в себя несколько этапов. В первую очередь выделяются специальные форматы записи, которые должны быть прочитаны определенным стандартизированным способом: телефон, дата, время, почтовый индекс и т.п. При этом анализируется вид записи (например, соответствует ли выражение стандартному виду записи типа XXX-XX-XX, XX:XX; входят ли цифры в возможный диапазон обозначения даты или времени, например, 13—30 или 60—65), а также наличие в предложении ключевых слов или словосочетаний (например, *телефон*, *мобильный*, *по московскому времени*…). Далее определяется разряд числительного (количественное или порядковое), прежде всего с помощью поиска ключевых слов, по преимуществу сочетающихся с порядковыми числительными (например, различные формы слова *год*). Следует заметить, что находящиеся в тексте римские цифры также должны быть расшифрованы как количественные или (чаще) порядковые числительные, например, *I квартал*, *Бенедикт XVI*.

Далее необходимо определить форму числительного, т.е. его падеж и (для числительных, обладающих данной категорией) род. При этом учитывается ближайший контекст числительного слева и справа: наличие предлога или другого управляющего слова слева (например, *к 23* „двадцати трем“, *до 23* „двадцати трех“, *владеет 23* „двадцатью тремя“ и т.п.), согласованного существительного или прилагательного справа (*10 пальцев* „десять“, *10 пальцами* „десятью“, *на 23 московских театральных площадках* „двадцати трех“, *10 этажа* „десятого“ и т.п.).

Снятие омонимии (омографии). Для синтеза речи наиболее важен анализ слов-омонимов, различающихся произношением (омографы), поскольку выбор между двумя омонимичными словоформами напрямую влияет на правильность синтезированного текста [5, 7]. Омографы в русском языке могут различаться ударением (например, *ст<u>о</u>ит—стои<u>т</u>*), а также наличием буквы „ё“, которая в современной орфографии чаще всего передается как *е* (*все—всё*), либо и тем и другим (*б<u>е</u>рег—бер<u>ё</u>г*).

Омонимичные словоформы могут иметь одинаковые грамматические признаки (например, *з<u>а</u>мок—зам<u>о</u>к*, *з<u>а</u>мка—замк<u>а</u>*…) либо различаться грамматическими характеристиками. В последнем случае омонимичными могут быть как различные словоформы внутри одной парадигмы (например, род.п. ед.ч.—им.п. мн.ч.: *облак<u>а</u>—<u>о</u>блака*, *стран<u>ы</u>—стр<u>а</u>ны*…), так и формы разных парадигм (например, существительное-инфинитив: *вест<u>и</u>—в<u>е</u>сти*, *проп<u>а</u>сть—пр<u>о</u>пасть*…). В случае с омонимами, одинаковыми по грамматическим характеристикам, разрешение неоднозначности может осуществляться только с помощью анализа лексического содержания предложения (ключевые слова, устойчивые выражения и т.п.). Если грамматические характеристики различаются, то можно использовать и анализ грамматического окружения слова для выбора омонима, подходящего по синтаксическому контексту. Усложняет проблему то, что омонимичные словоформы могут существенно различаться по частотности (на-

пример, *ух*а—*у*х*а*, *сорок*а—*сор*о*ка*, *кред*и*т*—*кр*е*дит*, *мо*ю—*м*о*ю*…). В таком случае зачастую становится продуктивным подход, при котором задаются специальные условия для нахождения низкочастотного омонимичного варианта, а в остальных случаях по умолчанию берется вариант с высокой частотностью.

Разрешение омонимии, как и расшифровка специальных обозначений, производится при помощи анализа контекста. На уровне индивидуальных слов-омонимов производится поиск в предложении ключевых слов или выражений. Этот этап включает анализ слов непосредственно рядом с текущим, как, например, в случае устойчивых выражений: *скрыто за семью замк*а*ми, в четырех стен*а*х*. Также анализируется состав предложения целиком, например: *Дверь была заперта на необычный зам*о*к* (ключевое слово *заперта*).

На уровне классов словоформ анализируется грамматическое окружение, т.е. выполняется поиск согласованных слов в предложении. Для формализации этого принципа были введены грамматические правила, увеличивающие условный „вес“ словоформы в зависимости от ее окружения. Правила хранятся в формализованном виде, позволяющем быстро оценивать и корректировать работу системы.

Результаты работы алгоритма лингвистического анализа. Для оценки качества лингвистического анализа в системе синтеза речи VitalVoice были проведены эксперименты по подсчету ошибок при обработке текстов. Для оценки правильности расшифровки нестандартных обозначений были взяты тексты с одного из новостных интернет-сайтов, поскольку данный тип текста содержит большое количество цифр, сокращений, специальных знаков и т.п. В ходе эксперимента был подсчитано число обозначений, неверно расшифрованных программой; результаты приведены ниже.

Расшифровка нестандартных обозначений

Слов в тексте, ед.	34235
Нестандартных обозначений в тексте, ед.	1066
Ошибок, ед	50
Ошибок, %	4,69
Правильно выполнено, %	95,31

Для оценки снятия омографии были взяты художественные тексты (произведения А.П.Чехова и Ю.В.Трифонова), поскольку они отличаются большим лексическим разнообразием. В ходе эксперимента был подсчитан процент слов-омографов, для которых было неверно определено место ударения; результаты приведены ниже.

Снятие омонимии (омографии)

Слов в тексте, ед.	37955
Омографов, ед.	2837
Ошибок, ед	113
Ошибок, %	3,98
Правильно выполнено, %	96,02

Обобщая результаты экспериментов, можно заметить, что лингвистический анализ, использующийся в системе VitalVoice, позволяет корректно разрешить неоднозначность чтения сложных элементов текста более чем в 95 % случаев. Основными источниками ошибок становятся сложные для анализа:

— случаи, когда для правильного прочтения элемента требуется анализ не только непосредственного контекста, но и дистанционных синтаксических связей. К примеру, во фрагменте: „*выбирать между 154 млрд кубометров по более низкой цене и 150 млрд по более высокой*“ второе числительное отделено несколькими другими членами предложения от относящегося к нему предлога;

— ошибочные или необщепринятые формы записи, например, *в 300-стах метрах* вместо *в 300-х метрах; 437 доллара* вместо *437 долларов*;

— формы записи, изначально не предназначенные для чтения вслух, такие как сложные цифровые записи, слова, полностью или частично замененные звездочками и т.п.

Развитие системы синтеза речи VitalVoice предполагает внедрение более глубокого синтаксического и семантического анализа текста, что позволит сократить количество ошибок, в особенности тех, которые связаны с недостаточно полным анализом предложения.

Заключение. Нормализация текста и определение правильного места ударения в слове — необходимый этап синтеза речи по тексту. Процедура морфологического и синтаксического анализа, реализованная в системе синтеза русской речи VitalVoice, позволяет выбрать корректный вариант прочтения таких элементов текста, как сокращения, цифры, специальные знаки, омографы и т.п. Как показывают эксперименты, проведенные на материале новостных и художественных текстов, точность правильного прочтения сложных элементов текста превышает 95 %.

СПИСОК ЛИТЕРАТУРЫ

1. *Taylor P.* Text to Speech Synthesis. Cambridge University Press, 2009.

2. *Sproat R.* et al. Normalization of Non-Standard Words // Computer Speech and Language. 2001. Vol. 15, N 3. P. 287—333.

3. *Allen J., Hunnicutt M. S., Klatt D*. From Text to Speech: The MI Talk system. Cambridge University Press, 1987.

4. *Liberman M. Y.* Text analysis and word pronunciation in text-to-speech synthesis // Advances in speech signal processing. 1992. P. 791—831.

5. *Иомдин Л. Л., Лобанов Б. М., Гецевич Ю. С.* Говорящий „ЭТАП“. Опыт использования синтаксического анализатора системы ЭТАП в русском речевом синтезе // Компьютерная лингвистика и интеллектуальные технологии: Матер. Междунар. конф. „Диалог“. М.: РГГУ, 2011. Вып. 10 (17). С. 669—679.

6. *Дружкин К. Ю., Цинман Л. Л.* Синтаксический анализатор лингвистического процессора ЭТАП-3: эксперименты по ранжированию синтаксических гипотез // Там же. М.: РГГУ, 2008. Вып. 7 (14). С. 147—153.

7. *Yarowsky D.* Homograph Disambiguation in Text-to-speech Synthesis // Progress in speech synthesis. 1996. P. 157—172.

Сведения об авторах

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

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

Илья Михайлович Аничкин — ООО „ЦРТ“, Санкт-Петербург; старший программист; E-mail: anichkin@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

СИСТЕМЫ РАСПОЗНАВАНИЯ ЛИЧНОСТЕЙ ПО ГОЛОСУ

УДК 004.93+57.087.1

Ю. Н. Матвеев

ИССЛЕДОВАНИЕ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ РЕЧИ ДЛЯ СИСТЕМ АВТОМАТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ДИКТОРОВ

Исследуется информативность речевых признаков наиболее популярных при создании автоматических систем идентификации дикторов. Эксперименты проводились на речевой базе данных, собранной в различных акустических условиях (широком диапазоне отношений сигнал/шум и уровней реверберации) и с использованием различных каналов записи.

Ключевые слова: *признаки речи, идентификация дикторов.*

Введение. Речевой сигнал существенно отличается от других акустических сигналов, так как произносится человеком для человека и служит для обмена информацией между людьми. Поэтому в системах распознавания личностей по голосу (распознавания дикторов) целью первичной обработки речевого сигнала является выделение признаков речи, специфичных для отдельных дикторов.

Наиболее распространенными речевыми признаками для систем идентификации дикторов являются [1]:

— частота основного тона;

— частота формант;

— кепстральные коэффициенты.

Первые два признака используются в основном в экспертных и полуавтоматических системах идентификации дикторов. В большинстве автоматических систем идентификации дикторов в качестве признаков используются векторы кепстральных коэффициентов:

— линейно-частотных кепстральных коэффициентов (LFCC, Linear-Frequency Cepstral Coefficients) или мэл-частотных кепстральных коэффициентов (MFCC, Mel-Frequency Cepstral Coefficients), получаемых по спектру Фурье [2];

— коэффициентов линейного предсказания (LPCC, Linear Prediction Cepstral Coefficients) [3];

— коэффициентов перцептивного линейного предсказания (PLP, Perceptual Linear Prediction) [4].

Наилучшим из критериев эффективности признаков является критерий разделимости классов, который связан с вероятностями ошибок классификатора. Поэтому для оценки информативности признаков или компактности пространств признаков распознаваемых голосов дикторов будет использоваться вероятностный критерий, связанный с величиной равновероятной ошибки (EER, Equal Error Rate), т.е. точкой равенства ошибок первого и второго рода, определяемой по пересечению кривых распределений вероятностей этих ошибок.

Значение EER характеризует в данном случае информативность признаков для текстонезависимой автоматической системы идентификации личности по речевому сигналу. Чем меньше значение EER, тем меньше перекрытие между кривыми ошибок первого и второго рода и тем компактнее пространства признаков.

Целью предлагаемой работы является оценка информативности различных кепстральных признаков для автоматической системы идентификации дикторов.

Оценка информативности речевых признаков на тестовой базе данных. С целью оценки информативности различных признаков для автоматической системы идентификации дикторов была использована речевая база данных [5], характеристики которой приведены в табл. 1.

Таблица 1

Параметр	Канал				
	1	2	3	4	5
Среднее значение ОСШ, дБ	35	20	40	8	4
Средний уровень реверберации, мс	250	300	200	650	1000
Количество фонограмм	377	548	398	352	817
Количество дикторов	76	123	72	80	105

В таблице 1 используются следующие обозначения каналов:

1) микрофонный канал (ближний микрофон — гарнитура), микрофон расположен на расстоянии не более 30 см от рта говорящего;

2) телефонный IP-канал;

3) телефонный GSM-канал;

4) микрофонный канал (удаленный микрофон), микрофон расположен на расстоянии 1—2 м от рта говорящего;

5) микрофонный канал (удаленный микрофон), микрофон расположен на расстоянии 2—4 м от рта говорящего.

Оценка информативности признаков проводилась с помощью автоматической системы идентификации дикторов, представленной на конкурс по распознаванию дикторов [6] NIST Speaker Recognition Evaluation (SRE) 2010, проведенный Институтом стандартов и технологий США (NIST).

В качестве исследуемых признаков были выбраны:

1) супервектор, составленный из 13 коэффициентов вектора MFCC, их 13 первых производных и их 13 вторых производных;

2) супервектор, составленный из 18 коэффициентов вектора LPCC и их 18 первых производных;

3) супервектор, составленный из 13 коэффициентов вектора PLP, их 13 первых производных и их 13 вторых производных.

В табл. 2 приведены результаты оценки информативности признаков на тестовой базе. Курсивом выделены минимальные значения EER, полужирным шрифтом — максимальные. Чем меньше значение EER, тем выше информативность признака.

Таблица 2

Признак	Канал				
	1	2	3	4	5
MFCC	4,0	*5,5*	**5,0**	10,0	21,5
LPCC	*3,0*	**8,5**	4,5	*6,0*	**26,5**
PLP	**5,0**	*5,5*	*3,5*	**12,0**	*17,5*

Анализ коррелированности признаков речевых признаков. Опыт участия в конкурсе NIST SRE-2010 [6] показал, что большинство мировых лидеров в своих системах используют не отдельные признаки, а их комбинации. При этом наблюдалось повышение эффек-

тивности идентификации даже при наличии корреляции между смешиваемыми признаками. Таким образом, при совместном использовании различных наборов признаков дополнительным критерием информативности признаков может быть степень их некоррелированности с другими признаками набора.

Так, в работе [7] отмечается коррелированность различных кепстральных признаков. Исследовались производные этих признаков (дельта-характеристики) для учета временных изменений. Включение производных в вектор признаков позволяет снизить влияние мультипликативных искажений сигнала, в силу того что эти искажения обычно медленно изменяются во времени и аддитивны в кепстральной области.

Из табл. 3 следует, что LPCC-коэффициенты имеют сильную корреляцию с MFCC-коэффициентами. Как отмечается в работе [7], это ожидаемый результат, поскольку оба этих признака описывают огибающую спектра. Кроме того, производные параметры кепстра также имеют высокую корреляцию, что объясняется схожестью методов их вычисления: ΔLPCC есть производная LPCC.

Таблица 3

Корреляция наборов признаков

Признак	ΔMFCC	LPCC	ΔLPCC
MFCC	0,77	0,88	0,71
ΔMFCC	—	0,73	0,69
LPCC	—	—	0,85

В работе [7] приведены результаты экспериментов по сравнению ряда других признаков, в том числе MFCC и PLP. Эксперименты проводились с использованием классификатора на основе смесей гауссовых распределений различного порядка (в зависимости от объема обучающих данных). Результаты исследований, приведенные в табл. 4, показали, что PLP не имеет преимуществ перед MFCC.

Таблица 4

Надежность идентификации
(в процентах правильно идентифицированных дикторов)

Порядок модели	MFCC	PLP
2	95,36	82,26
4	97,14	93,93
8	98,33	96,79
16	99,52	98,10
32	99,05	98,45

В обзоре [8] сделан вывод о том, что различные кепстральные признаки, такие как MFCC, LFCC, LPCC и PLP, имеют сильную корреляцию. Однако возможно их комбинирование (смешивание) для повышения надежности идентификации [7].

В табл. 5 дана оценка средней корреляции (СКО = 0,01) признаков по каналам 1—4 тестовой базы данных (см. табл. 1). Наиболее коррелированными признаками снова оказались MFCC и LPCC, а наименее — LPCC и PLP. Полученное значение корреляции признаков MFCC и LPCC согласуется с полученным в работе [7] и приведенным в табл. 3.

Таблица 5

Признак	LPCC	PLP
MFCC	0,84	0,81
LPCC		0,69

В табл. 6 дана оценка средней корреляции (СКО = 0,01) признаков по каналу 5 тестовой базы данных (см. табл. 1). Данный канал характеризуется высоким уровнем реверберации (более 1000 мс) и низким соотношением сигнал-шум (4 дБ). В таких акустических условиях наиболее коррелированными оказались признаки MFCC и PLP, а наименее — LPCC и PLP.

Таблица 6

Признак	LPCC	PLP
MFCC	0,70	0,82
LPCC	—	0,57

В табл. 7 приведены результаты экспериментов по комбинированию признаков, которые согласуются с приведенными выше оценками.

Таблица 7

Признак	Канал									
	1		2		3		4		5	
	Вес	EER, %	Вес	EER, %	Вес	EER, %	Вес	EER, %	Вес	EER, %
MFCC	0,004	4,0	0,465	**5,5**	0,034	10,0	0,240	5,0	0,174	21,5
LPCC	0,790	**3,0**	0,005	8,5	0,766	**6,0**	0,220	4,5	0,136	26,0
PLP	0,206	5,0	0,530	**5,5**	0,200	12,0	0,542	**3,5**	0,690	**17,5**
Комбинация (смесь)	1	2,5	1	4,5	1	6,0	1	3,0	1	17,0

Из полученных результатов можно сделать следующие выводы:

1) комбинирование (смешивание) признаков всегда обеспечивает наименьшее значение EER;

2) признак, имеющий наименьшее значение EER, всегда имеет наибольший весовой коэффициент (вес);

3) признак PLP менее коррелирован с MFCC и LPCC, чем MFCC и LPCC между собой, поэтому он всегда имеет значимый вес;

4) признаки MFCC и LPCC имеют высокую степень корреляции, поэтому один из них часто вносит очень мало дополнительной информации в обобщенное решение.

Заключение. В настоящей работе исследована информативность широко известных наборов речевых признаков, таких как MFCC, LFCC, LPCC и PLP. В качестве критерия информативности для отбора признаков в системе идентификации дикторов по голосу использовалось значение EER.

Показано, что MFCC, LPCC и PLP имеют сильную корреляцию, а также, что ни один из рассмотренных признаков не дает преимуществ по сравнению с другими по уровню информативности в различных акустических условиях и в различных каналах записи. Однако возможно их комбинирование для повышения надежности идентификации дикторов по голосу. Результат смешивания признаков всегда обеспечивает наименьшее значение EER.

СПИСОК ЛИТЕРАТУРЫ

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

2. *Huang X., Acero A., Hon H.* Spoken Language Processing: A guide to theory, algorithm, and system development. Prentice Hall, 2001. 1008 p.

3. *Zheng F., Zhang G., Song Z.* Comparison of Different Implementations of MFCC // J. Computer Sci. and Techn. 2001. Vol. 16, N 6. P. 582—589.

4. *Hermansky H., Malayath N.* Speaker Verification Using Speaker-Specific Mappings // Proc. of the Workshop on Speaker Recognition and its Commercial and Forensic Applications. Avignon, 1998. P. 111—114.

5. База данных для идентификации говорящего по голосу “RUASTEN”. Регистрационное свидетельство № 2010620533 от 20.09.2010.

6. *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.

7. *He W., Hong P.* The Application of Fusion Technology for Speaker Recognition // Intern. J. of Computer Science and Network Security. 2007. Vol. 7, N 12. P. 300—303.

8. *Kinnunen T., Li H.* An overview of text-independent speaker recognition: From features to supervectors // Speech Communication. 2010. Vol. 52, N 1. P. 12—40.

Сведения об авторе

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

Рекомендована кафедрой
речевых информационных систем

Поступила в редакцию
22.10.12 г.

УДК 681.3

Т. С. Пеховский, А. Ю. Сизов

СРАВНЕНИЕ РАЗЛИЧНЫХ СМЕСЕЙ ГАУССОВЫХ PLDA-МОДЕЛЕЙ В ЗАДАЧЕ ТЕКСТОНЕЗАВИСИМОГО РАСПОЗНАВАНИЯ ДИКТОРА

Исследуется актуальность использования классической смеси PLDA-моделей с распределением Гаусса в качестве априорного в пространстве i-векторов для задачи верификации диктора. Исследуются условия эксперимента, в которых это использование выгодно при существующих ограничениях размеров обучающих баз. Показано, что в рамках кроссканальной задачи использование смеси двух PLDA-моделей эффективнее, чем традиционная схема с использованием одной PLDA-модели.

Ключевые слова: *i-вектор, совместный факторный анализ, смесь PLDA-моделей, распознавание диктора.*

Введение. В последнее десятилетие активно развиваются технологии текстонезависимого распознавания личностей по голосу (дикторов). В работах Рейнольдса впервые было предложено для таких задач использовать смеси гауссовых распределений (Gaussian Mixture Models, GMM) [1, 2]. В работе [2] была показана эффективность универсальной фоновой модели (Universal Background Model, UBM), также показана эффективность MAP-адаптации (Maximum A-Posteriori Probability) модели GMM-UBM при получении модели диктора.

Модель GMM-UBM обычно обучается на большой базе дикторов, с использованием критерия максимального правдоподобия и, как правило, имеет 2048 компонент. Модель диктора здесь получается путем адаптации только средних модели GMM-UBM и последующей конкатенации отдельных компонент, с формированием при этом GMM-супервектора средних — высокоразмерного вектора признаков $m(s, h)$ для h-й сессии s-го диктора.

Работы Кенни [3—5] посвящены модели совместного факторного анализа (Joint Factor Analysis, JFA) и ее различным редуцированным версиям [6—8]. JFA — это порождающая модель, используемая с целью эффективного решения проблем междикторской и межсессионной вариативности диктора в GMM-подходе. Модель JFA можно использовать (см., например, [9]) для получения оценок верификации по критерию Неймана—Пирсона. Прогресс

современных систем верификации диктора обусловлен использованием новых низкоразмерных векторов признаков, порождаемых одной из версий JFA. В этой новой модели [10] не выполняется расщепление пространства GMM-супервектора на дикторское и канальное подпространства. Процесс обучения *T*-матрицы полной изменчивости [10] аналогичен процессу обучения матрицы собственных голосов [3], за исключением того, что

— в случае матрицы собственных голосов все сессии обучающего диктора конкатенируются для последующего обучения;

— в случае *T*-матрицы все сессии обучающего диктора расцениваются как произведенные различными дикторами.

Таким образом, вектор полной изменчивости $w(s, h)$ [10] сохраняет зависимость и от канала, и от диктора и является полным низкоразмерным аналогом супервектора $m(s, h)$. Задача расщепления пространства полной изменчивости на подпространство диктора и подпространство канала реализуется, например, с помощью линейного дискриминантного анализа (Linear Discriminate Analysis, LDA). Дальнейшее развитие текстонезависимого распознавания диктора связано большей частью с использованием векторов $w(s, h)$ в качестве входных векторов-признаков — i-векторов.

Результаты последних конкурсов по оцениванию систем распознавания дикторов (Speaker Recognition Evaluation, SRE) Национального института стандартов и технологий (National institute of Standards and Technologies, NIST) [11] показали высокую эффективность различных методов, использующих низкоразмерные i-векторы. Среди них самыми перспективными являются методы, основанные на модели вероятностного линейного дискриминантного анализа (Probabilistic LDA, PLDA) [12, 13]. В работе [12], посвященной распознаванию лиц, было представлено точное решение процедуры обучения гауссовой PLDA-модели (G-PLDA) с использованием критерия максимального правдоподобия. В работе [13] Кенни реализовал вариационное байесовское обучение PLDA-модели для верификации диктора с использованием тяжелохвостых распределений (HT-PLDA), отметив, что *t*-распределение Стьюдента должно более адекватно описывать такие негауссовы эффекты канала, как грубые искажения речи в случае записи через удаленный микрофон. Модель HT-PLDA продемонстрировала высокую эффективность при тестировании на однородном телефонном корпусе. Дальнейшее развитие подхода PLDA показало, что такую же эффективность систем верификации можно получить при использовании G-PLDA-модели, если осуществить нормализацию длины i-вектора [14].

В настоящей работе исследуются условия, при которых актуально использование классических смесей моделей G-PLDA [12], обучаемых „без учителя“ (unsupervised mixtures, U-mix) в пространстве i-векторов. U-mix позволяют осуществлять нелинейное покрытие структуры плотности данных обучающей базы, не требуя исходного знания о сегментации данных, что должно повысить эффективность системы верификации на тестовой базе, имеющей подобную структуру. По мнению авторов настоящей статьи, применение U-mix PLDA будет более актуальным в той ситуации, когда в обучающей базе априори существуют физически разнородные кластеры. Примером такой постановки задачи может являться стандартная для NIST кроссканальная задача верификации диктора, в которой обучающая база содержит данные, полученные в микрофонных и телефонных каналах.

Следует отметить, что работа [15] посвящена использованию смесей PLDA для решения кроссгендерной задачи верификации. Но, в отличие от предлагаемой нами U-mix-системы, в работе [15] обучались отдельные PLDA-системы для двух полов (компоненты смеси), обучаемые „с учителем“ (supervised mixtures, S-mix), на сегментированном материале своих полов, а смесь PLDA-моделей была реализована путем мягкого байесовского комбинирования достоверностей отдельных PLDA-систем.

В настоящей работе также ставится цель сравнить эффективность систем верификации диктора, построенных на базе моделей U-mix PLDA и на базе S-mix PLDA-моделей по схеме Кенни [16].

Обучение моделей U-mix PLDA. Поскольку в работе [12] формулы обновления гиперпараметров для G-PLDA-модели представлены без вывода, детально опишем точный вывод процедуры обучения смеси на основе критерия максимального правдоподобия.

Модель G-PLDA. Каждая из компонент рассматриваемой смеси PLDA-моделей состоит из единственной гауссовой модели фактора диктора, определенного в пространстве i-векторов. Формальное отличие от классического факторного анализа (Factor Analysis, FA) [17] заключается в том, что обучающий *s*-й диктор представлен своими *R*(*s*) сессиями, что, в свою очередь, характерно для схемы обучения PLDA-модели:

$$\begin{pmatrix} D^{(s,1)} \\ \vdots \\ D^{(s,R(s))} \end{pmatrix} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} + \begin{bmatrix} U & \cdots & 0 & V \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & U & V \end{bmatrix} \begin{pmatrix} x^{(s,1)} \\ \vdots \\ x^{(s,R(s))} \\ y^{(s)} \end{pmatrix} + \begin{pmatrix} \varepsilon^{(s,1)} \\ \vdots \\ \varepsilon^{(s,R(s))} \end{pmatrix} = \underline{D}^{(s)} = \underline{\mu} + \underline{A}\underline{z}^{(s)} + \underline{\varepsilon}^{(s)}, \qquad (1)$$

где μ — F-мерный вектор средних; $V = (F \times Q_y)$-матрица, столбцы которой можно трактовать как собственные голоса; $U = (F \times Q_x)$-матрица, ее столбцы — это собственные каналы, а шумовая (F×F)-матрица ковариации Σ — общая для всех моделей в смеси. Легко заметить, что для каждой *r*-й сессии (1) приобретает вид:

$$D^{(s,r)} = \mu + \begin{bmatrix} U & V \end{bmatrix} \begin{pmatrix} x^{(s,r)} \\ y^{(s)} \end{pmatrix} + \varepsilon^{(s,r)} = \mu + Wh^{(s,r)} + \varepsilon^{(s,r)}.$$

Здесь y, x, $\varepsilon^{(s,r)} \propto N(0,\Sigma)$ — скрытые переменные, представляющие факторы диктора, факторы канала и шум соответственно. Будем предполагать гауссов характер априорных распределений этих переменных.

Построение смеси G-PLDA моделей. Начинаем с построения функции правдоподобия смеси PLDA, состоящей из *M* моделей, используя обучающую базу из независимых дикторов, имеющих по *R*(*s*) сессий. Тогда логарифм функции правдоподобия на неполных данных есть:

$$L = \sum_{s}^{S} \ln \left\{ \sum_{m}^{M} \pi_m p_m(\underline{D}^{(s)} \mid \theta_m) \right\},$$

где π_m — веса смеси, $\theta_m = \{W_m, \mu_m, \Sigma\}$ — гиперпараметры *m*-й модели, а маргинальное правдоподобие $p_m(D^{(s)} \mid \theta_m)$ относится к отдельной вероятностной модели PLDA и выражается как

$$p_m(\underline{D}^{(s)} \mid \theta_m) = \int p_m(\underline{D}^{(s)} \mid \theta_m, \underline{z}_m) p(\underline{z}_m) d\underline{z}_m.$$

Здесь с вектором данных *s*-го диктора $\underline{D}^{(s)}$ связывается ряд бинарных скрытых переменных $\rho_m^{(s)} \in \{0,1\}$, $\sum_{m=1}^{M} \rho_m^{(s)} = 1$. Тогда параметры для этой модели смеси могут быть определены стандартным EM-алгоритмом [17] с использованием функции правдоподобия на полных данных L_c:

$$L_c = \sum_s^S \sum_m^M \rho_m^{(s)} \ln\left\{\pi_m p_m(\underline{D}^{(s)}, \underline{z}_m^{(s)} \mid \theta_m)\right\}, \tag{2}$$

где совместная вероятность:

$$p_m(\underline{D}^{(s)}, \underline{z}_m^{(s)} \mid \theta_m) = p_m(\underline{D}^{(s)} \mid \theta_m, \underline{z}_m^{(s)}) p(\underline{z}_m^{(s)}) =$$
$$= (2\pi)^{-R(s)F/2} \mid \underline{\Sigma} \mid^{-1/2} \exp\left\{-\tfrac{1}{2}(a_m^{(s)})^T \underline{\Sigma}^{-1}(a_m^{(s)})\right\} (2\pi)^{-Q/2} \exp\left\{-\tfrac{1}{2}\underline{z}_m^{(s)T} \underline{z}_m^{(s)}\right\}. \tag{3}$$

В формуле (2) $Q = Q_y + R(s)Q_x$, а $a_m^{(s)}$ есть вектор:

$$a_m^{(s)} = \left(\underline{D}^{(s)} - \underline{A}_m \underline{z}_m^{(s)} - \underline{\mu}_m\right).$$

Далее, следуя модели смеси FA [17], для математического ожидания полной функции $< L_c >$ относительно апостериорного распределения $P(z|D)$, легко получить:

$$< L_c >= \sum_s^S \sum_m^M \gamma_m(s)\Big[\ln \pi_m - \frac{1}{2} < \underline{z}_m^{(s)T} \underline{z}_m^{(s)} > - \frac{1}{2}\ln \mid \underline{\Sigma} \mid -$$
$$-\frac{1}{2}\Big\{(\underline{D}^{(s)} - \underline{\mu}_m)^T \underline{\Sigma}^{-1}(\underline{D}^{(s)} - \underline{\mu}_m) - 2(\underline{D}^{(s)} - \underline{\mu}_m)^T \underline{\Sigma}^{-1} \underline{A}_m < \underline{z}_m^{(s)} > +$$
$$+\mathrm{tr}[\underline{A}_m{}^T \underline{\Sigma}^{-1} \underline{A}_m < \underline{z}_m^{(s)} \underline{z}_m^{(s)T} >]\Big\}\Big] + \mathrm{const}.$$

Перейдем от схемы полного вектора z к представлению вектора h. Этот переход весьма облегчает последующие формулы обновления параметров в М-шаге EM-алгоритма и является очевидным, если рассмотреть скаляр под знаком экспоненты в формуле (3):

$$< (a_m^{(s)})^T \underline{\Sigma}^{-1}(a_m^{(s)}) >_{P(z|D)} = < \sum_{r=1}^{R(s)} (\xi_m^{(s,r)})^T \hat{\Sigma}^{-1}(\xi_m^{(s,r)}) >_{P(h|D)},$$

где $\xi_m^{(s,r)}$ есть вектор:

$$\xi_m^{(s,r)} = \left(D^{(s,r)} - W_m h_m^{(s,r)} - \mu_m\right).$$

Тогда математическое ожидание полной функции $< L_c >$ относительно апостериорного распределения $P(z|D)$ будет иметь вид:

$$< L_c >= \sum_s^S \sum_m^M \gamma_m(s)\Big[\ln \pi_m - \frac{R(s)}{2}\ln \mid \Sigma \mid -$$
$$-\frac{1}{2}\Big\{\sum_r^{R(s)} (D^{(s,r)} - \mu_m)^T \Sigma^{-1}(D^{(s,r)} - \mu_m) - 2\sum_r^{R(s)} (D^{(s,r)} - \mu_m)^T \Sigma^{-1} W_m < h_m^{(s,r)} > +$$
$$+\mathrm{tr}\Big[\sum_{r=1}^{R(s)} W_m{}^T \Sigma^{-1} W_m < h_m^{(s,r)} h_m^{(s,r)T} >\Big]\Big\}\Big] + \mathrm{const},$$

где компоненты парного вектора h и его ковариации должны браться из компонент полного вектора z и его ковариации [12]:

$$< h_m^{(s,r)} > \leftarrow < z_m^{(s,r)} >,$$
$$< h_m^{(s,r)} h_m^{(s,r)T} > \leftarrow < z_m^{(s,r)} z_m^{(s,r)T} >,$$

найденных, как будет описано далее, на Е-шаге EM-алгоритма. Тогда на М-шаге, в стационарной точке для функции $< L_C >$, будем иметь следующие формулы для обновления параметров:

$$\pi_m = \frac{N_m}{N} = \frac{1}{\sum_s^S \sum_m^M \gamma_m^{(s)}} \sum_s^S \gamma_m^{(s)}, \quad \mu_m = \frac{\sum_s^S \gamma_m^{(s)} \sum_r^{R(s)} (D^{(s,r)} - W_m < h_m^{(s,r)} >)}{\sum_s^S \gamma_m^{(s)} R(s)},$$

$$W_m = \left[\sum_s^S \gamma_m^{(s)} \sum_r^{R(s)} (D^{(s,r)} - \mu_m) < h_m^{(s,r)} >^T \right] \left[\sum_s^S \gamma_m^{(s)} \sum_r^{R(s)} < h_m^{(s,r)} h_m^{(s,r)T} > \right]^{-1}, \qquad (4)$$

$$\Sigma = \frac{\operatorname{diag}\left[\sum_s^S \sum_m^M \gamma_m^{(s)} \sum_r^{R(s)} < \left(D^{(s,r)} - W_m h_m^{(s,r)} - \mu_m\right)\left(D^{(s,r)} - W_m h_m^{(s,r)} - \mu_m\right)^T > \right]}{\sum_s^S \sum_m^M \gamma_m^{(s)} R(s)}.$$

Заметим, что в настоящей работе везде используется шумовая матрица ковариации Σ — общая для всех анализаторов. В формуле (4) представлен ее диагональный случай. E-шаг EM-алгоритма для смеси PLDA-моделей стандартен, так как он будет выполнен в представлении полного вектора z. На этом шаге [17] необходимо найти апостериорное распределение

$$< \underline{z}_m^{(s)} >= \underline{\Sigma}_m^{(Z)} \, \underline{A}_m{}^T \, \underline{\Sigma}^{-1} (\underline{D}(s) - \underline{\mu}_m)$$

и соответствующую матрицу:

$$< \underline{z}_m^{(s)} \, \underline{z}_m^{(s)T} >= \underline{\Sigma}_m^{(z)} + < \underline{z}_m^{(s)} >< \underline{z}_m^{(s)T} >,$$

где апостериорная матрица ковариации для обобщенного скрытого вектора z есть

$$\underline{\Sigma}_m^{(z)} = (\underline{I} + \underline{A}_m{}^T \, \underline{\Sigma}^{-1} \, \underline{A}_m)^{-1},$$

I — единичная матрица.

Также необходимо найти $\gamma_m^{(s)}$ (responsibilities) — апостериорное распределение для набора скрытых переменных $\rho_m^{(s)}$, обслуживающих смесь [17]:

$$\gamma_m^{(s)} = \frac{\rho_m^{(s)}}{\sum_k^M \rho_k^{(s)}} = \frac{\pi_m p_m(\underline{D}^{(s)})}{\sum_k^M \pi_k p_k(\underline{D}^{(s)})} = \frac{\pi_m \int p_m(\underline{D}^{(s)} \mid z) p(z) dz}{\sum_k^M \pi_k \int p_k(\underline{D}^{(s)} \mid z) p(z) dz},$$

находим точное значение маргинального правдоподобия (evidence):

$$p_m(\underline{D}^{(s)}) = \int p_m(\underline{D}^{(s)}, z) dz = \int p_m(\underline{D}^{(s)} \mid z) p(z) dz =$$

$$= (2\pi)^{-FR(s)/2} \mid \underline{C}_m \mid^{-1/2} \exp\left\{ -\tfrac{1}{2} (\underline{D}^{(s)} - \underline{\mu}_m)^T \, \underline{C}_m^{-1} (\underline{D}^{(s)} - \underline{\mu}_m) \right\} \qquad (5)$$

(здесь и далее для удобства записи будем опускать θ_m).

И, таким образом, выражение для логарифма ответственностей:

$$\ln \rho_m^{(s)} = \ln(\pi_m) - \frac{1}{2} \ln \left| \underline{C}_m \right| - \frac{1}{2} (\underline{D}^{(s)} - \underline{\mu}_m)^T \, \underline{C}_m^{-1} (\underline{D}^{(s)} - \underline{\mu}_m) + \text{const},$$

где матрица ковариации $\underline{C}_m$ в (5), после взятия интеграла для вектора диктора $\underline{D}^{(s)}$, состоящего из $R(s)$ сессий, может быть представлена как:

$$\underline{C}_m = \underline{\Sigma} + \underline{A}_m \underline{A}_m{}^T =$$

$$= \begin{bmatrix} \Sigma & & & \\ & \Sigma & & \\ & & \ddots & \\ & & & \Sigma \end{bmatrix} + \begin{bmatrix} U_m U_m{}^T + V_m V_m{}^T & V_m V_m{}^T & \cdots & V_m V_m{}^T \\ V_m V_m{}^T & U_m U_m{}^T + V_m V_m{}^T & \cdots & \vdots \\ \vdots & \vdots & \ddots & V_m V_m{}^T \\ V_m V_m{}^T & \cdots & V_m V_m{}^T & U_m U_m{}^T + V_m V_m{}^T \end{bmatrix}.$$

Обращение матриц ковариации $\underline{C}_m$ и $\underline{\Sigma}_m^{(z)}$ представляет при точном выводе определенную трудность. Но их обращение может быть сведено к обращению отдельных блоков.

Стадия верификации. *Случай U-mix PLDA*. Оценка PLDA для смеси имеет ту же структуру, что и оценка для отдельной PLDA-модели [13]:

$$\text{Score} = \ln \frac{P(D_1, D_2 \mid T)}{P(D_1 \mid I) P(D_2 \mid I)},$$

где выражение для маргинального правдоподобия в числителе (случай $R(s)$=2) и двух — в знаменателе (случай $R(s)$=1) посчитано, в отличие от [13], точно:

$$P(\underline{D}^{(s)}) = \sum_m^M \pi_m \int p_m(\underline{D}^{(s)} \mid z) p(z) dz =$$

$$= \sum_m^M \pi_m \left[(2\pi)^{-R(s)F/2} \mid \underline{C}_m \mid^{-1/2} \exp\{-\tfrac{1}{2} (\underline{D}^{(s)} - \underline{\mu}_m)^T \underline{C}_m^{-1} (\underline{D}^{(s)} - \underline{\mu}_m)\} \right]$$

и, согласно (1), представляет собой достоверность смеси PLDA-моделей.

Случай S-mix PLDA. Представим реализацию S-mix PLDA по Кенни [16], состоящую из M отдельных PLDA-моделей:

$$\text{Score} = \ln \frac{P(D_1, D_2 \mid T)}{P(D_1, D_2 \mid I)} = \ln \frac{\sum_m P(D_1, D_2 \mid m, T) P(m \mid T)}{\sum_{m,m'} P(D_1 \mid m, I) P(m \mid I) P(D_2 \mid m', I) P(m' \mid I)} =$$

$$= \ln \frac{\sum_m P(D_1, D_2 \mid m, T) P(m \mid T)}{\sum_{m,m'} Q^{(m,m')} P(D_1 \mid m, I) P(D_2 \mid m', I)},$$

где априорные распределения для целевых дикторов и „самозванцев“ (imposters) выбираются равными для каждой m-й компоненты смеси Кенни [16]:

$$P(m \mid T) = P(m \mid I) = 1 / M,$$

$$Q^{(m,m')} = P(m \mid I) P(m' \mid I) = 1 / M^2.$$

Таким образом, это можно рассматривать как вариант байесовского комбинирования отдельных PLDA-систем на стадии верификации.

Эксперименты. *Предобработка речевого сигнала.* Все записи были сегментированы на участки „речь“ и „пауза“. Участки „пауза“ затем были удалены из записей. В экспериментах использовались 39-мерные мэл-частотные кепстральные коэффициенты (mel-frequency cepstral coefficients, MFCC) [1]. MFCC-векторы состояли из 13 кепстральных коэффициентов, их первых и вторых производных, вычисляемых по 5 соседним кадрам. Использовались кадры с окном 22 мс и со сдвигом окна в 11 мс. Каждый кадр был преэмфазирован [1] и домножен на окно Хэмминга. Также везде применялась стандартная процедура вычитания кепстрального среднего из кепстральных коэффициентов.

Универсальная фоновая модель (UBM). Использовалась гендернезависимая UBM, имеющая 512 компонент и полученная с помощью EM-обучения на основе критерия максимального правдоподобия на телефонных базах NIST SRE 1998—2008 годов (все языки, оба пола). Системы PLDA обучались на записях голосов 4329 мужчин и женщин. Использовалась диагональная, а не полноковариационная GMM-UBM.

Кроссканальный экстрактор i-векторов. В кроссканальной задаче необходимо использовать универсальный экстрактор i-векторов, который бы мог адекватно работать как в телефонном, так микрофонном каналах. Здесь проблемой является несбалансированность количества записей в телефонном и микрофонном каналах. Последних в несколько раз меньше в базах NIST, чем первых. В этом случае, как предложено в работе [18], используется универсальный экстрактор i-векторов, который бы подходил как для микрофонных записей речи, так и для телефонных. Он основан на отдельных оценках максимального правдоподобия двух T-матриц полной изменчивости. Математически это можно выразить для дикторо- и каналозависимого супервектора μ следующим образом:

$$\mu = \mu_0 + T'w' + T''w''. \qquad (6)$$

В настоящей работе телефонная T' матрица с 400 базисными столбцами обучена на 11 256 телефонных записях из NIST 2002/2003/2004/2005/2006/2008 от 1250 дикторов-мужчин (только английский язык). Микрофонная T'' матрица той же размерности обучалась на 4705 микрофонных записях из NIST 2005/2006/2008 от 203 дикторов-мужчин (только английский язык), согласно [18]. Таким образом была решена проблема значительной несбалансированности наборов телефонных и микрофонных записей. После оценки T'' и T' конкатенируются, чтобы получить смешанную T-матрицу:

$$\mu = \mu_0 + Tw, \qquad (7)$$

где w-векторы есть интересующие нас итоговые i-векторы. Таким образом, используется кроссканальный экстрактор i-векторов размерности с 700 базисными столбцами.

Однородный экстрактор i-векторов. В кроссканальной задаче также будет использоваться обычный экстрактор i-векторов (6), но обученный только на телефонных записях, назовем его однородным экстрактором i-векторов. Такой необычный, на первый взгляд, выбор объясняется следующими причинами. Апостериорное распределение i-векторов обучающей базы экстрактора i-векторов (7), согласно JFA, всегда будет близко к его априорному $N(0,1)$. Таким же распределение i-векторов будет и для любой другой базы, близкой по условиям записи к обучающей (по каналу, по полу, по языку и т.д.). Но, как показали эксперименты, при существенном рассогласовании базы обучения и тестовой базы всегда наблюдается существенный сдвиг центра распределения i-векторов тестовой базы относительно нуля. Это приводит к деградации равновероятной ошибки первого и второго рода (Equal Error Rate, EER) системы, основанной на одной PLDA-модели. Но для случая обучения, например, двух PLDA моделей на двух физически явных кластерах (например, каналы в кроссканальной задаче) такое поведение однородного экстрактора будет способствовать разделению кластеров в пространстве i-векторов. Идея заключается в том, что таким образом улучшаются условия применения смеси PLDA-моделей в пространстве i-векторов, которое изначально более подходит под одну модель. Кроме того, будет использоваться однородный телефонный экстрактор i-векторов T'.

Переход в LDA-пространство. Как уже было отмечено выше, JFA-экстрактор i-векторов генерирует i-векторы, содержащие информацию как о дикторе, так и о канале. Поэтому еще одним условием, способствующим успешному применению смеси PLDA, будет переход от входных i-векторов к их проекциям, получаемым в результате LDA-преобразования. Это позволяет:

— уменьшить канальный шум;

— получить добавочную редукцию размерности входных векторов.

Такая верификационная схема $\mathrm{TV} \to \mathrm{LDA} \to \mathrm{PLDA}$ была успешно применена в различных работах по верификации диктора, а именно в кроссгендерных [15] и кроссканальных [16, 19] задачах. Метод LDA широко используется для редукции размерности в задачах классификации. В нашей работе LDA-преобразование редуцирует i-векторы до 200-мерного пространства, заполненного собственными векторами, соответствующими самым большим собственным значениям следующей обобщенной задачи о собственных значениях λ и собственных векторах x:

$$S_b x = \lambda S_w x, \tag{8}$$

где S_b и S_w — соответственно матрицы межклассовой и внутриклассовой вариативности. После решения обобщенной задачи (8) получаем LDA-матрицу, которую применяем к i-векторам в обучающих и тестовых базах. Были построены две LDA-матрицы. В случае кроссканального экстрактора обучалась LDA-матрица размерностью 700×200 на данных обучения этого экстрактора, в случае однородного экстрактора — LDA-матрица размерностью 400×200 только на 11 256 телефонных записях, использованных для обучения однородного экстрактора.

LDA-проекции i-векторов затем подвергались процедуре нормализации, согласно [14], но только для тестовой базы (U-L-G конфигурация в терминах [14]). Эта нормализация состоит в проектировании LDA-векторов на единичную сферу.

Условия обучения. Обучались две модели S-mix G-PLDA (M=2, 3) и две U-mix G-PLDA (M=1, 2). Для модели S-mix PLDA (M=3) независимо были обучены (везде — только английский язык):

— Phone-PLDA — модель, обученная на 11 256 телефонных записях из NIST 2002/2003/2004/2005/2006/2008 от 1250 дикторов-мужчин;

— Mic-PLDA — модель, обученная на 4705 микрофонных записях из NIST 2005/2006/2008 от 203 дикторов-мужчин;

— CI-PLDA — каналонезависимая PLDA-модель, обученная на совокупном наборе данных систем Phone-PLDA и Mic-PLDA.

При обучении возникает проблема сильной несбалансированности наборов телефонных и микрофонных записей NIST. Авторы решили эту проблему, взяв из 11 256 только 5000 телефонных записей дикторов, которые были представлены в микрофонном канале, и добавив к этому набору все записи по микрофонному каналу. Так же, как и в работе [16], модель S-mix PLDA (M=3) выполнена с помощью комбинирования этих трех моделей на стадии получения оценок, а S-mix PLDA (M=2) состояла из комбинации двух систем — Phone-PLDA и Mic-PLDA. Обучение компонент проводилось согласно вариационному байесовскому выводу Кенни [13]. Модели U-mix PLDA (M=1, 2) обучались на всем смешанном наборе данных двух систем Phone-PLDA и Mic-PLDA. Везде количество столбцов матрицы собственных голосов V для всех PLDA-моделей было $Q_y = 200$, а U=0. Везде в целях ускорения сходимости при обучении на основании максимального правдоподобия добавлялись итерации минимизации дивергенции Кульбака—Лейблера фазы обучения по Кенни [13]. Шумовая матрица ковариации Σ в (4) для всех случаев имела полноковариационный вид.

Результаты тестирования для кроссканала (det3). Результаты сравнения моделей U-mix и S-mix PLDA относительно результатов основного (core-core) теста на мужских голосах базы NIST SRE 2010 для кроссканальной задачи (det3) [11] представлены в табл. 1. Для оценки эффективности систем использовались ошибка EER и новый нормализованный минимум функции стоимости обнаружения NIST (Minimum Detection Cost Function, minDCF) как метрика [11].

Таблица 1

Система	*M*=1	*M*=2	*M*=3
S-mix G-PLDA Кроссканальный экстрактор	—	4,31 % [0,598]	3,83 % [0,577]
U-mix G-PLDA Кроссканальный экстрактор	3,82 % [0,579]	3,70 % [0,535]	—
U-mix G-PLDA Однородный экстрактор	4,06 % [0,601]	3,22 % [0,525]	—

Из табл. 1 следует, во-первых, что модель S-mix G-PLDA лучше всего работает при *M*=3 и осуществляет относительную редукцию EER системы на 11 % при *M*=2, а во-вторых, что модель U-mix G-PLDA при *M*=2 немного выигрывает (EER=3,70 %) у лучшей S-mix-системы при *M*=3 (EER=3,83 %) даже при использовании кроссканального экстрактора. Наконец, лучшей (EER=3,22 %) оказалась модель S-mix G-PLDA при *M*=2, использующая однородный экстрактор.

Результаты тестирования для телефонного канала (det5). Результаты сравнения систем верификации, полученных на неконтролируемой смеси PLDA-моделей, для однородного (телефон) по каналу условия (det5) представлены в табл. 2. Целью эксперимента было выяснить, можно ли наблюдать на однородном корпусе (телефон, мужчины, английский язык) структуру плотности, соответствующую выбору более чем одной модели G-PLDA. Из табл. 2 видно, что S-mix G-PLDA при *M*=2 существенно проигрывает (EER=3,97 %) системе G-PLDA (EER=3,69 %).

Таблица 2

Система	*M*=1	*M*=2
U-mix G-PLDA Однородный экстрактор	3,69 % [0,532]	3,97 % [0,585]

Обсуждение. Как ожидалось, идея однородного экстрактора оказалась весьма полезной для использования моделей U-mix PLDA. Однородный экстрактор породил на тестовой базе det3 такую же двухкластерную (телефон—микрофон) структуру плотности в пространстве i-векторов, что и в обучающем множестве. Это непосредственно следует из сравнения 2-й и 3-й строк табл. 1, видно, что в случае U-mix G-PLDA при *M*=2 во время обучения на основе максимального правдоподобия произошел захват смесью этой структуры, что положительно повлияло на эффективность этой системы (EER=3,22 %) и негативно — на эффективность системы на основе модели U-mix G-PLDA при *M*=1 (EER возрос с 3,82 до 4,06 %). Последнее свидетельствует о несоответствии структуры данных, порожденной однородным экстрактором, модели одной G-PLDA. Напротив, как следует из табл. 2, в случае однородного тестового условия (det5) эта структура, порожденная однородным экстрактором, соответствует одной модели G-PLDA. Можно сказать, что на текущий момент количество дикторов в доступных речевых базах недостаточно для эффективного использования смесей PLDA-моделей при *M*>1 в случае однородной базы данных. Таким образом, проведенные тестовые эксперименты показывают эффективность подхода моделей U-mix PLDA для кроссканальной задачи верификации диктора, которая превосходит по эффективности модель S-mix G-PLDA [16].

Заключение. В статье предложено использовать модель U-mix PLDA для решения кроссканальной задачи верификации диктора. Проведенные эксперименты на данных NIST SRE 2010 позволяют сделать следующие выводы.

1. На однородных базах данных использовать более одной модели нецелесообразно, даже в пространстве LDA-векторов, так как существующие обучающие базы на данный момент не обладают достаточным количеством дикторов.

2. На кроссканальной задаче смеси PLDA моделей можно успешно применять, но в пространстве LDA-векторов и при использовании однородного экстрактора.

3. Схема однородного экстрактора в совокупности со смесью двух моделей оказывает существенную конкуренцию схеме кроссканального экстрактора с одним гауссовым анализатором в стандартной кроссканальной задаче NIST.

В будущем планируется реализовать модель U-mix G-PLDA при использовании полной байесовской структуры. Это позволит автоматически определять релевантную размерность матриц факторов диктора и канала, а также количество компонент смеси для обучающей базы.

Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации.

СПИСОК ЛИТЕРАТУРЫ

1. *Reynolds D. A., Rose R. C.* Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE Trans. Speech Audio Process. 1995. N 3. P. 72—83.

2. *Reynolds D. A., Quatieri T. F., Dunn R. B.* Speaker Verification Using Adapted Gaussian Mixture Models // Digit. Signal Process. 2000. N 10. P. 19—41.

3. *Kenny P.* Joint factor analysis of speaker and session variability: Theory and algorithms // Technical report CRIM-06/08-13. 2005.

4. *Kenny P., Boulianne G., Ouellet P., Dumouchel P.* Joint factor analysis versus eigenchannels in speaker recognition // IEEE Trans. Audio, Speech, Lang. Process. 2007. Vol. 15. P. 1435—1447.

5. *Kenny P., Ouellet P., Dehak N., Gupta V., Dumouchel P.* A Study of Inter-Speaker Variability in Speaker Verification // IEEE Trans. Audio, Speech and Lang. Process. 2008. Vol. 16. P. 980—988.

6. *Vogt R., Sridharan S.* Explicit modeling of session variability for speaker verification // Comput. Speech and Lang. 2008. Vol. 22. P. 17—38.

7. *Burget L., Matejka P., Glembek O., Cernocky J.* Analysis of feature extraction and channel compensation in GMM speaker recognition system // IEEE Trans. on Audio, Speech and Lang. Process. 2007. Vol. 15. P. 1979—1986.

8. *Pekhovsky T., Oparin I.* Eigen Channel Method for Text-Independent Russian Speaker Verification // Proc. of the XII Intern. Conf. “Speech and Comput.” SpeCom'08. Moscow, Russia, 2008. P. 385—390.

9. *Glembek O., Burget L., Brummer N., Kenny P.* Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis // IEEE Int. Conf. on Acoust., Speech, and Signal Process. Taipei, Taiwan, 2009.

10. *Dehak N., Kenny P., Dehak R., Dumouchel P., Ouellet P.* Front-end factor analysis for speaker verification // IEEE Trans. on Audio, Speech, and Lang. Process. 2010. Vol. 19. P. 788—798.

11. [Электронный ресурс]: <http://www.itl.nist.gov/iad/mig/tests/sre>.

12. *Prince S. J. D., Elder J. H.* Probabilistic linear discriminant analysis for inferences about identity // Proc. 11th Intern. Conf. on Comput. Vision. Rio de Janeiro, Brazil, 2007. P. 1—8.

13. *Kenny P.* Bayesian speaker verification with heavy tailed priors // Proc. Odyssey Speak. and Lang. Recognit. Workshop. Brno, Czech Republic, 2010.

14. *Garcia-Romero D., Espy-Wilso C. Y.* Analysis of i-vector length normalization in speaker recognition systems // Proc. of Interspeech. Florence, Italy, 2011. P. 249—252.

15. *Senoussaoui M., Kenny P., Brummer N., Villiers E., Dumouchel P.* Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition // Proc. of Interspeech. Florence, Italy, 2011. P. 25—28.

16. *Simonchik K., Pekhovsky T., Shulipa A., Afanasev A.* Supervised Mixture of PLDA Models for Cross-Channel Speaker Verification // Proc. Interspeech. Portland, USA, 2012.

17. *Tipping M., Bishop C. M.* Mixtures of probabilistic principal component analyzers // Neural Comput. 1999. Vol. 11. P. 443—482.

18. *Senoussaoui M., Kenny P., Dehak N., Dumouchel P.* An i-vector extractor suitable for speaker recognition with both microphone and telephone speech // Proc. Odyssey Speak. Recognit. Workshop. Brno, Czech Republic, 2010.

19. *Senoussaoui M., Kenny P., Dumouchel P., Castaldo F.* Well-calibrated heavy tailed Bayesian speaker verification for microphone speech // Proc. ICASSP. Prague, Czech Republic, 2011.

Сведения об авторах

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; доцент; E-mail: tim@speechpro.com

Александр Юрьевич Сизов — студент; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра речевых информационных систем; E-mail: sizov@speechpro.com

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 004.934.2

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

КЛАССИФИКАЦИЯ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ДИКТОРА С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОПОРНЫХ ВЕКТОРОВ И КРИТЕРИЯ ДЖИНИ

Исследована эффективность применения критерия Джини для формирования пространства признаков SVM-классификатора. Приведены результаты экспериментального определения оптимального набора информативных признаков и построения классификатора.

Ключевые слова: *речь, классификация эмоционального состояния, критерий Джини, метод опорных векторов, автоматический выбор информативных признаков.*

Введение. Исследование паралингвистических средств речевой коммуникации включает определение довольно разнообразных характеристик: эмоциональное состояние, пол и возраст диктора, стиль разговора, уровень заинтересованности, сонливость и даже наличие алкогольного опьянения.

В настоящей работе исследуется задача определения эмоционального состояния говорящего человека (диктора). При решении этой задачи возникает ряд трудностей [1]: отсутствует четкое определение эмоции, отсутствует однозначный ответ на вопрос о соотнесении акустических особенностей речи диктора с его эмоциональным состоянием. Все это приводит к различиям в формах классификации эмоций и произвольной расстановке акцентов разными группами исследователей [2].

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

1) вычисление базовых характеристик речевого сигнала (low-level descriptors, согласно терминологии [4]); оценка мощности, частоты основного тона F_0 (ЧОТ), формантных частот, спектральных и кепстральных характеристик речевого сигнала и т.д.;

2) вычисление функционалов от базовых характеристик, таких как перцентили, экстремумы и их отношения, моменты высших порядков, коэффициенты регрессии и т.д.;

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

В настоящей работе предложено использовать статистический критерий, отражающий сходство видов распределений исследуемой характеристики при решении задачи классификации эмоциональных состояний.

Описание базы тестирования. Обучение и тестирование алгоритма проводилось на записях, взятых из Берлинской базы данных эмоциональной речи (Emo-DB) [6]. Данная база была собрана в Техническом университете Берлина и неоднократно использовалась исследователями при разработке систем распознавания эмоционального состояния. Исследование базы показало [6], что эмоции в ней распознаются слушателями в 80 % случаев, и в 60 % признаются естественными.

Методология классификации эмоционального состояния диктора. Обобщенная структурная схема системы определения эмоционального состояния диктора приведена на рис. 1.

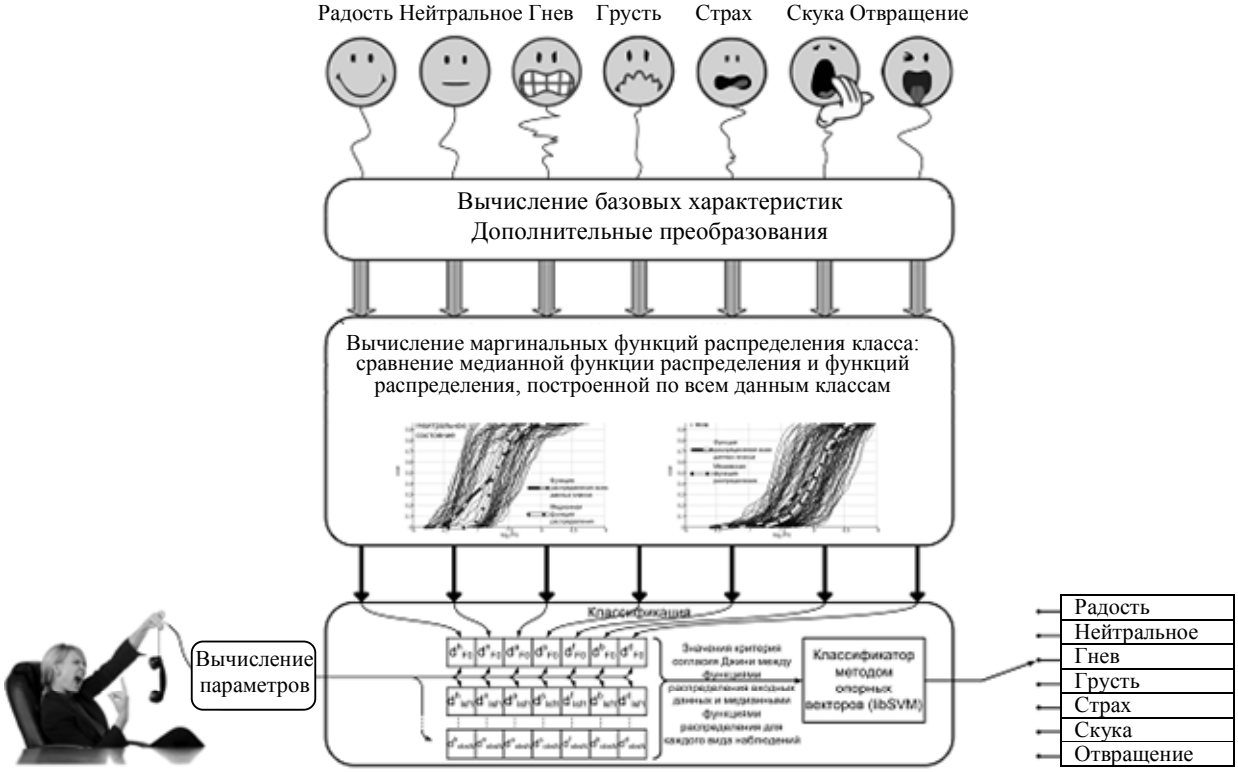


Рис. 1

Предобработка заключается в умножении каждой записи на случайный коэффициент усиления от –20 до +20 дБ, чтобы исключить привязку к абсолютному уровню сигнала, предыскажению и пропусканию сигнала через полосовой фильтр с полосой пропускания от 300 до 3400 Гц.

Вычисление параметров. Исследования в области психологии и психолингвистики предоставили сведения о множестве акустических, просодических и лингвистических характеристик речи, способных служить информативными признаками при распознавании эмоционального состояния и проявляющихся на уровне голосовых сегментов, слогов и целых слов. При этом во множестве видов характеристик выделяют базовые и вычисленные из них функционалы. Набор базовых характеристик в нашем случае включает: кратковременную оценку мощности сигнала; оценку частоты основного тона в соответствии с алгоритмом, рассмотренным в [7]; джиттер (модуляция частоты основного тона) и шиммер (модуляция амплитуды сигнала); коэффициенты линейных спектральных частот; кепстральные коэффициенты, вычисленные на основе коэффициентов линейного предсказания; фонетическую функцию на основе вычисления лог-спектрального расстояния, расстояния Итакуры-Саито и COSH-расстояния [8]; коэффициенты вещественного кепстра; мел-кепстральные коэффициенты; оценки асимметрии и эксцесса распределения ошибки линейного предсказания сигнала [9]; энергетический оператор Тигера в формантных полосах и критических полосах слуха [10]; отношения мощностей в формантных полосах.

К вычисленным базовым признакам применялся ряд преобразований: вычисление первой и второй производных, применение энергетического оператора Тигера, вычитание медианного значения, стандартизация.

Вычисление статистического критерия. Наиболее простым решением является вычисление статистических критериев (расстояния) между двумя многомерными плотностями распределения. Однако для построения многомерных плотностей распределения требуется большое количество обучающих данных. В противном случае классификатор может оказаться неустойчивым. Поэтому было решено вместо многомерных использовать маргинальные распределения каждой исчисленной характеристики сигнала.

Выбор критерия вычисления расстояния между двумя функциями распределения играет важную роль. Для построения пространства расстояний необходимо использовать функцию расстояния, не требующую априорных предположений о видах распределения сравниваемых величин. Поэтому в качестве функции расстояния было решено использовать критерий Джини [11]:

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

Таким образом, для каждого исследуемого признака обрабатываемого файла необходимо вычислить критерий Джини между распределением этого признака для анализируемого сигнала и распределением, описывающим каждый класс эмоциональной речи. Эти расстояния целесообразно использовать как пространство наблюдений для обучения и тестирования классификатора. При этом для описания функций распределения класса подходящим представляется использование функции распределения всех данных этого класса. Однако при этом способе формирования функции не учитывается поведение каждой эмпирической функции распределения, входящей в этот класс, и таким образом теряется информация совокупности эмпирических функций распределения класса как семейства некоторых кривых.

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

$$\tilde{F}(x) = \underset{i}{\text{median}} \left(F_i(x) \right).$$

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

На рис. 2 приведен пример различных способов формирования функции распределения класса: как функции распределения всех наблюдений (*1*, Class CDF) и как медианной функции распределения (*2*, Median CDF).

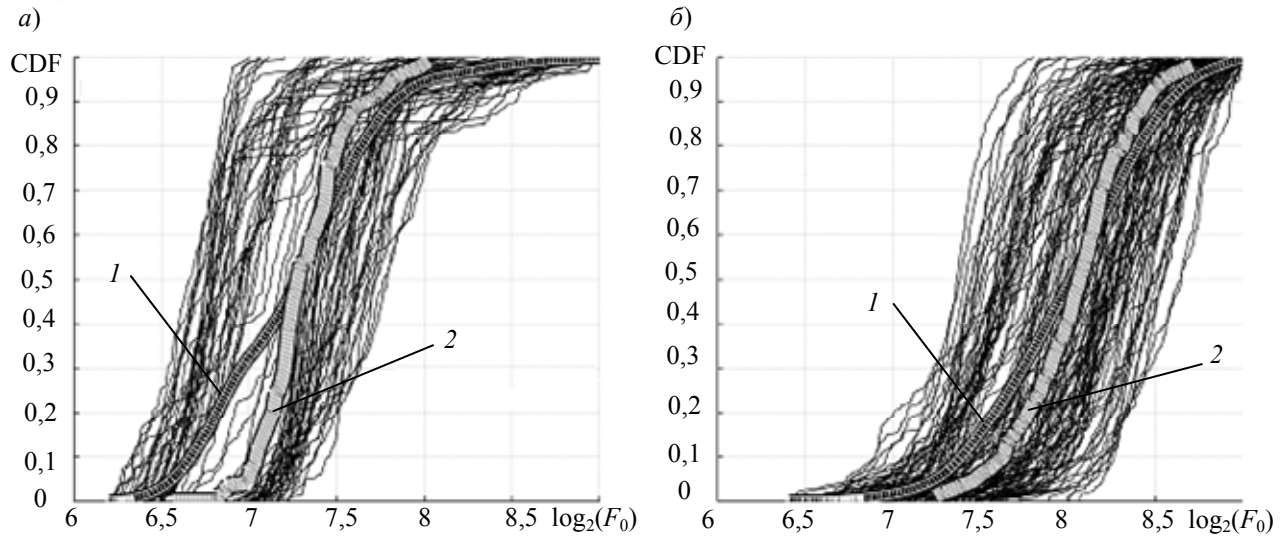


Рис. 2

На рис. 2, *а* (нейтральное эмоциональное состояние) хорошо заметно наличие двух областей группирования функций распределения, связанных с различием распределений ЧОТ мужских и женских голосов (график на рис. 2, *б* соответствует состоянию гнева). При этом

следует отметить, что Median CDF, в отличие от Class CDF, в значительной степени сохраняет форму поведения кривых функций распределения класса.

Классификация объектов в сформированном пространстве признаков является завершающей операцией большинства систем определения эмоционального состояния диктора.

Для выполнения классификации был использован метод опорных векторов (SVM-метод), реализованный в библиотеке libSVM [12]. В данной библиотеке мультиклассовый SVM-классификатор строится как набор классификаторов „каждый-с-каждым“ с последующим голосованием. Это позволяет на этапе определения оптимального набора признаков выбрать лучший набор именно для мультиклассовой классификации. Для построения разделяющей гиперповерхности использовалось RBF-ядро [12] как наиболее универсальное и не требующее априорных предположений о характере распределения наблюдений. Эффективность распознавания в процессе определения оптимального набора информативных признаков и подбора параметров модели оценивалась при помощи метода K-кратной кросспроверки, реализованного в составе пакета libSVM.

Эксперимент. Рассмотрим результаты экспериментального исследования описанного способа формирования пространства признаков для построения классификатора. Как указывалось выше, экспериментальные исследования проводились с использованием базы Emo-DB. При этом для оценки эффективности классификации данных использовалось среднее значение диагональных элементов нормированной матрицы неточностей (average recall).

Автоматическое определение оптимального набора информативных параметров. Для определения этого набора использовался алгоритм последовательного выбора параметров (Sequential Feature Selection, SFS), нашедший широкое применение при решении задач распознавания эмоциональных состояний по голосу [3]. Суть его работы заключается в том, что на каждой итерации к набору добавляются признаки, обеспечивающие наибольший прирост эффективности классификации. Чтобы увеличить гибкость алгоритма, на каждой его итерации после добавления некоего, обеспечивающего максимальный прирост эффективности классификации, множества из m признаков, производится удаление множества из n признаков. Нами использовались значения m=5 и n=3.

Зависимость эффективности распознавания P эмоциональных состояний от числа добавленных в набор алгоритмом SFS информативных признаков N показана на рис. 3.

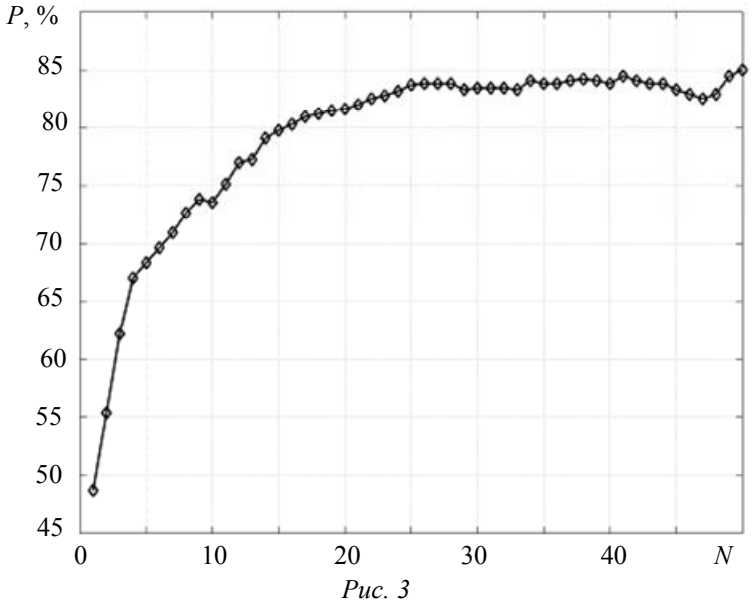


Рис. 3

Из рисунка видно, что на каждой итерации эффективность распознавания эмоциональных классов возрастала до тех пор, пока не достигла некоторого максимума, соответствую-

щего примерно N=25. Дальнейшее повышение эффективности распознавания с ростом количества информативных признаков происходило очень медленно.

Классификация. Оценка эффективности классификации проводилась при помощи K-кратной кросспроверки, при этом K = 10 [13]. В таблице приведена усредненная матрица неточностей, полученная для 25 информативных признаков. Средняя эффективность, достигнутая построенным классификатором, оказалась ≈ 83 %.

Фактический класс	Предсказанный класс						
	гнев	скука	отвращение	страх	радость	нейтральное	грусть
Гнев	**91,7**	0	0	1,9	6,4	0	0
Скука	0	**90,3**	0	0	0	6,6	3,1
Отвращение	1,8	6,7	**68,6**	6,3	4,1	3,9	8,6
Страх	5,5	2,0	1,4	**74,6**	7,9	2,1	6,4
Радость	20,2	0	0	6,3	**73,5**	0	0
Нейтральное	0	9,3	0	0	0	**88,3**	2,5
Грусть	0	4,0	0	0	0	3,3	**92,7**

Выводы и направления дальнейших исследований. В статье предложен новый метод формирования пространства признаков классификатора на основе вычисления критерия Джини. Было проведено экспериментальное исследование эффективности метода, включающее этапы определения оптимального набора параметров и построения SVM-классификатора. Экспериментальное исследование проводилось на базе Emo-DB [6]. В качестве показателя эффективности использовалось среднее значение диагональных элементов нормированной матрицы неточностей, а для оценки точности прогнозирования — метод K-кратной кросспроверки.

В качестве дальнейшей работы представляется целесообразным протестировать эффективность применения описанного метода для классификации других паралингвистических средств речевой коммуникации.

СПИСОК ЛИТЕРАТУРЫ

1. *El Ayadi M., Kamel M.S., Karray F.* Survey on speech emotion recognition: Features, classification schemes, and databases // Pattern Recognition. 2011. Vol. 44, N 3. P. 572—587.

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

3. *Schuller B., Batliner A., Steidl S., Seppi D.* Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge // Speech Communication. 2011. Vol. 53, N 9—10. P. 1062—1087.

4. *Eyben F., Wöllmer M., Schuller B.* OpenEAR-Introducing the Munich open-source emotion and affect recognition toolkit // Proc. 3rd Intern. Conf. on Affective Computing and Intelligent Interaction. ACII. 2009. P. 1—6.

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

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

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

8. *Rabiner L. R., Biing-Hwang Juang.* Fundamentals of speech recognition. Prentice Hall, 1993. 507 p.

9. *Nemer E., Goubran R., Mahmoud S.* Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain // IEEE Transactions on Speech and Audio Processing. 2001. Vol. 9, N 3. P. 217—231.

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

11. *Кобзарь А. И.* Прикладная математическая статистика. М.: ФИЗМАТЛИТ, 2006. 816 с.

12. *Chang C.-C., Lin C.-J.* LIBSVM: a library for support vector machines // ACM Transactions on Intelligent Systems and Technology. 2011. Vol. 2, N 27. P. 1—27.

13. *Kohavi R.* A study of cross-validation and bootstrap for accuracy estimation and model selection // Proc. of the 14th Intern. Joint Conf. on Artificial Intelligence. 1995. Vol. 2. P. 1137—1143.

Сведения об авторах

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

Андрей Геннадьевич Давыдов — канд. техн. наук; ООО „Речевые технологии“, Минск; старший научный сотрудник; E-mail: davydov-a@speechpro.com

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 004.93+57.087.1

Д. В. Дырмовский, С. Л. Коваль

ОСОБЕННОСТИ ЧЕЛОВЕКО-МАШИННОГО ИНТЕРФЕЙСА СОВРЕМЕННЫХ СИСТЕМ БИОМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ

Обоснованы требования к организации человеко-машинного интерфейса для современных систем автоматической и автоматизированной идентификации личности, основанных на анализе биометрических признаков.

Ключевые слова: *человеко-машинный интерфейс, идентификация личности, биометрическая система идентификации, голосовая биометрия.*

Введение. Системы автоматической идентификации личности (САИ) по биометрическим признакам получают все большее распространение. Они предназначены для решения задач учета и мониторинга неизвестных лиц, выполнения криминалистических идентификационных экспертиз. Рассмотрим особенности интерфейса и структуры САИ на примере использования динамических идентификационных признаков речевого сигнала.

Оптимальное представление результатов работы САИ. В современных САИ используется процедура обучения на больших базах биометрических данных совпадающих и различающихся личностей. Например, в обучающих базах речевых САИ содержатся файлы речи нескольких тысяч дикторов, записанных в разных условиях. Корпусной подход позволяет автоматически выбрать оптимальные правила и параметры идентификации, сопоставляющие скалярному расстоянию x между сравниваемыми речевыми файлами значения вероятности совпадения/различия дикторов. САИ сравнивает пары дикторов и по найденному для них x выдает вероятностный результат тождества/различия дикторов, практическая интерпре-

тация которого может различаться для разных задач. Например, для задачи верификации достаточно принимать решение „Да“—„Нет“.

Для решения задачи идентификации необходимы оценка точности предлагаемого решения и оценка его неопределенности. Характеристики речевого сигнала существенно зависят от свойств каналов звукозаписи и звукопередачи, состояния диктора, типа речевой коммуникации, сопутствующих помех и искажений и т.п. В силу этого создать представительные обучающие базы данных невозможно. Существующие учетные САИ [1—3] выдают результат только в виде ранжированного списка сравниваемых дикторов, что неприемлемо для единичных сравнений. Интерфейс САИ должен позволять пользователю выбрать оптимальную форму представления результатов и учесть характеристики конкретных сравниваемых дикторов, ориентируясь на случаи и множественного, и единичного сравнения. Ни одна из существующих прикладных и исследовательских САИ этим требованиям не удовлетворяет [1—6].

Известно много способов представления результатов работы САИ [7—9], однако выбор способа, оптимального для речевых систем, не очевиден. САИ проводит поиск целевого диктора в списке проверяемых. Предлагается представлять результат работы САИ в виде списка похожих дикторов (СП), полученного усечением списка всех проверяемых дикторов на основе предлагаемых показателей оценки работы системы. Будем исходить из того, что любая САИ для каждой пары целевой диктор—проверяемый диктор вычисляет наборы характеристических признаков и скалярное расстояние x между этими наборами. Используются показатели: $FRR(x)$ — False Rejection Rate — оценка вероятности ошибки 1-го рода: вероятность того, что файлы с речью целевого („своего“) диктора из списка проверяемых файлов не попадут в СП, если расстояние x от них до эталона будет больше соответствующего заданному значению FRR. $FAR(x)$ — False Acceptance Rate — оценка вероятности ошибки 2-го рода: вероятность того, что файлы с речью нецелевого („чужого“) диктора из списка проверяемых файлов попадут в СП, если расстояние x от них до эталона будет меньше соответствующего заданному значению FAR. $LR(x)$ — likelihood Ratio — оценка отношения правдоподобия для гипотез совпадения и различия дикторов при данном x между сравниваемыми файлами. $LR(x)$ рассчитывается как отношение вероятности отклонить „своего“ диктора при расстоянии между сравниваемыми дикторами больше данного x к вероятности принять „чужого“ диктора за своего при расстоянии между сравниваемыми файлами меньше x:

$$LR(x) = \frac{FRR(x)/\big(FRR(x)+FAR(x)\big)}{FAR(x)/\big(FRR(x)+FAR(x)\big)} = \frac{FRR(x)}{FAR(x)}. \qquad (1)$$

Формула (1) имеет следующее толкование: вероятность верности нулевой гипотезы для интервала значений расстояния между дикторами больше данного x равна отношению доли попавших в этот интервал совпавших пар дикторов (т.е. $FRR(x)$) к общему числу пар дикторов, расстояние между которыми попало в этот интервал (т.е. $FRR(x)+FAR(x)$). Аналогично толкуется и знаменатель дроби в формуле.

Введем следующие понятия: P — общая вероятность совпадения сравниваемых дикторов; DET-график (Detection Error Trade-off) — график зависимости FRR от FAR.

Пример представления этих величин для системы автоматической идентификации VoiceNet приведен в таблице и на рис. 1 и 2. На рис. 1 представлены результаты идентификации САИ VoiceNet одного диктора (его данные заданы строкой в верхнем окне экрана) при сравнении с большим списком дикторов. Результат каждого сравнения указан в строках нижнего окна экрана. В каждой строке указан номер сравниваемого диктора в списке сравнения, идентификатор в базе данных, FRR, FAR, LR, имя секции базы данных, имя карточки диктора, имя звукового файла, тип источника звука. На рис. 2 приведен DET-график для результата сравнения двух дикторов в системе VoiceNet.

FRR, %	*FAR*, %	*LR*, у.е.	*P*
29,07	0,01	2907	0,999
7,20	0,10	72	0,986
1,07	0,50	2,1	0,68
1,00	0,56	1,78	0,64
0,67	1,00	0,67	0,40
0,53	5,00	0,11	0,1
0,50	6,74	0,07	0,07
0,10	33,44	0,003	0,003

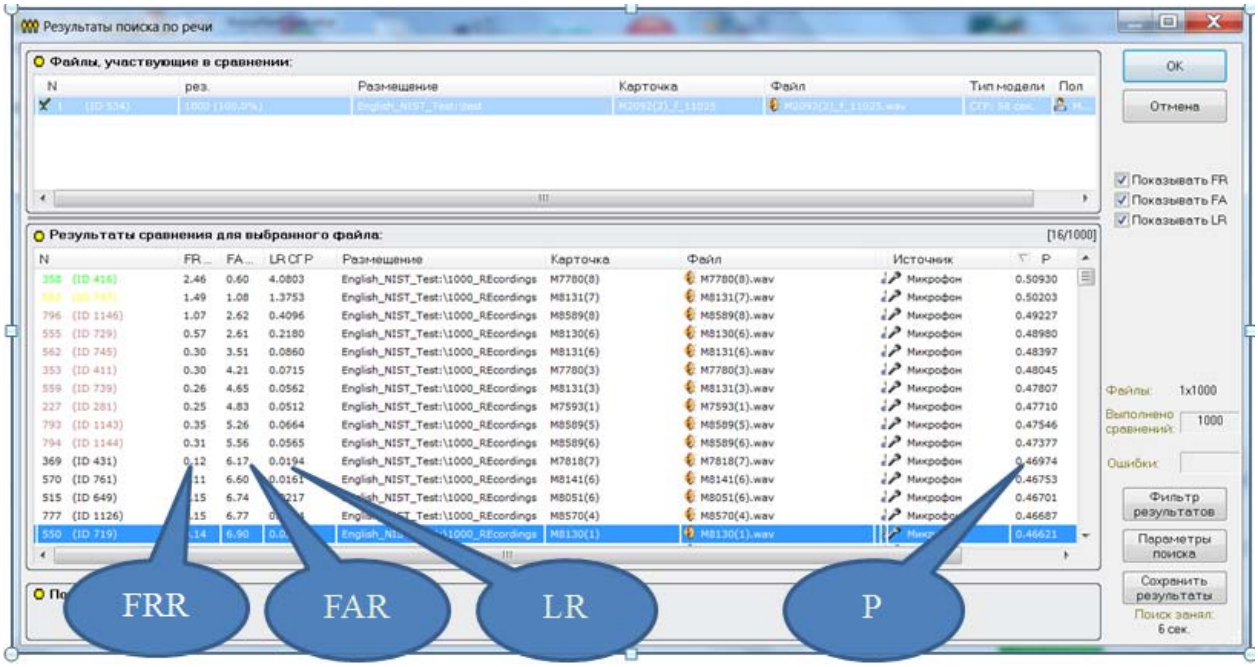


Рис 1

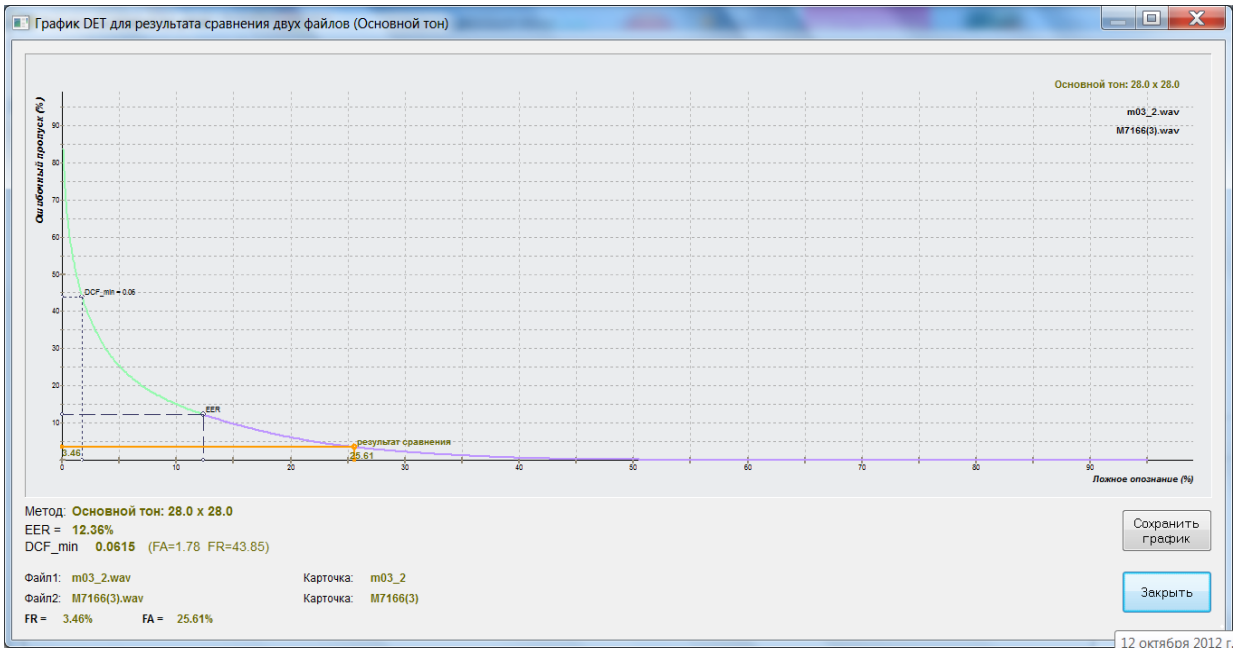


Рис. 2

Проиллюстрируем возможность использования этих показателей для конкретной САИ [10]. После сравнения диктора с данными базы из 100 000 голосов получим СП, ранжированный по степени сходства с данным диктором. Если ограничить рассматриваемый СП только дикторами, для которых $LR > 2907$, то диктор, находящийся в списке проверяемых, попадет в СП с вероятностью ≈70 %, а с вероятностью ≈30 % — не попадет. Такой способ выбора порога обладает существенным преимуществом. Соответствующее выбранному порогу значение $FAR = 0{,}01$ % означает, что в СП попадет не более 10 „чужих“ дикторов. Пользователь САИ за относительно малое время может проверить „ручными“, трудоемкими, экспертными средствами реальность тождества целевого диктора с этими 10 дикторами и выяснить, есть он действительно среди них или нет. Для 10 дикторов такая проверка на практике реализуема, а для большего числа дикторов уже трудноосуществима. Такой порог выгодно применять при проверке по большой базе неизвестных дикторов. Пропуск „своего“ диктора вероятен, но ввиду ограниченности ресурсов операторов системы по „ручной“ проверке СП выбор такого порога часто является единственной возможностью обнаружить в базе неизвестного искомого диктора.

При ограничении рассматриваемого СП только дикторами, для которых $LR > 0{,}67$, если целевой диктор есть в списке проверяемых, то он не попадет в СП с вероятностью всего ≈0,7 %. Однако у такого выбора порога отсечки есть существенный недостаток. Соответствующее выбранному порогу значение $FAR = 1$ % означает, что наиболее вероятно в СП попадет около 1000 „чужих“ дикторов, которые САИ сочтет близкими к целевому. Проверить „ручными“ средствами реальность тождества этих 1000 дикторов с целевым затруднительно. Тем не менее, такой порог выгодно применять при проверке по малой базе проверяемых дикторов. DET-график дает возможность выбора подходящих порогов отсечки для конкретной задачи с еще большей точностью, чем таблица.

Применение концепции доверительности данных к работе САИ. САИ применяются и в судебных экспертизах [4, 5, 11—14] при сравнении всего двух объектов. В этом случае результаты работы САИ целесообразно применять в рамках так называемого байесовского подхода [8, 15—17], объединяя данные исследований различных методов в единой формуле на основе значений *LR* по каждому из методов. Однако возможность применения статистических результатов обучения САИ к единичному случаю совершенно неочевидна. САИ вычисляет *LR*, что требует оценки неопределенности измерения [18]. Значение *P* можно обоснованно считать оценкой случайной величины, а в качестве параметров неопределенности результата оценки предлагается считать границы односторонних доверительных интервалов (ДИ) [19], которые определяются на основе подхода, близкого к методике NIST [20], которая использовалась для сравнительной оценки неопределенности результатов различных САИ. Нами рассматривается оценка неопределенности результатов работы отдельной САИ.

На этапе обучения САИ получает распределения частоты встречаемости расстояний x для пар совпадающих и различающихся дикторов. При решении задачи идентификации возможны два варианта: H_0 — сравниваемые дикторы совпадают и H_1 — различаются. Пусть $P(H_0|x)$ — апостериорная вероятность правильности гипотезы о совпадении дикторов. Тогда, согласно формуле Байеса:

$$P(H_0|x) = \frac{P(x|H_0)P(H_0)}{P(x|H_0)P(H_0) + P(x|H_1)P(H_1)}, \qquad (2)$$

где $P(H_0)$ и $P(H_1)$ — априорные вероятности гипотез, $P(x|H_0)$ и $P(x|H_1)$ — вероятности получения x при верности каждой из гипотез. Значения $P(H_0)$ и $P(H_1)$ для простоты полагаются равными.

Апостериорная вероятность $P(H_0|x)$ моделируется сигмоидной функцией зависимости от x [21], оценка параметров которой проводится на обучающей базе данных. На рис. 3

приведен пример зависимости апостериорной вероятности $P(H_0|x)$ от x, полученной для конкретной САИ [10] на основе анализа гистограмм распределений x для совпадающих (эллипсы) и различающихся (крестики) дикторов в обучающей базе данных.

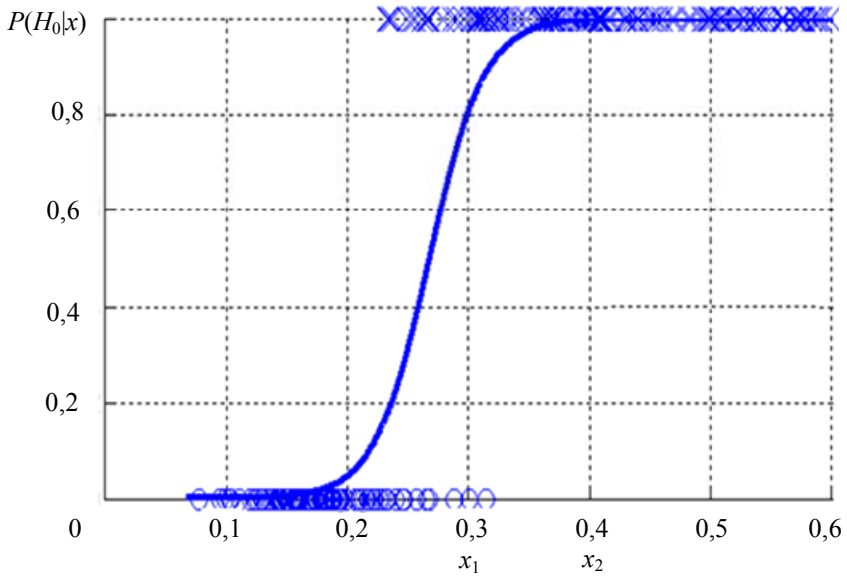


Рис. 3

Однако реальная форма распределения x для больших обучающих баз данных речевых САИ обычно далека от какого-либо стандартного типа. В силу этого при оценке ДИ для P предлагается использовать непараметрический bootstrap-метод, не требующий предположений о форме оцениваемого распределения [7, 20, 22]. Он позволяет оценивать распределение P для генеральной совокупности, используя только одну большую выборку. Пусть оценка апостериорной вероятности $\hat{P}_n = \hat{P}_n(H_0|x)$ получена для начальной выборки $(X_1, ..., X_n)$. Чтобы оценить ДИ, в который с заданной вероятностью попадают значения $P(H_0|x)$, конструируем из начальной выборки большое количество других выборок, выбирая в произвольном порядке ее элементы „с возвратом“. Создадим наборы из B возможно повторяющихся элементов $(X_1^*, ..., X_n^*)$ и вычислим для них соответствующие значения $\hat{P}_n^*(b)$, $b = 1, ..., B$. На основе оценок для bootstrap-выборок вычислим bootstrap-распределение $\hat{P}_n^*$: $G_*(p) = P\{\hat{P}_n^* \le p\}$ — аналог распределения наборов обычных выборок из генеральной совокупности. Соответствующие процентили этого распределения определяют квантили уровня значимости α и $1-\alpha$ $G_*^{-1}(\alpha) = \inf\{x : G_*(x) \ge \alpha\}$ и $G_*^{-1}(1-\alpha)$ как нижнюю и верхнюю границы $1-2\alpha$ ДИ для оценки $\hat{P}_n = \hat{P}_n(H_0|x)$ [22].

В качестве характеристик САИ выберем односторонний доверительный интервал (ОДИ). Верхний ОДИ:

$$P(-\infty < P(H_0|x) \le \hat{P}_n^{\,u}) = \alpha, \tag{3}$$

где $\hat{P}_n^{\,u} = G_*^{-1}(1-\alpha)$, и нижний ОДИ:

$$P(\hat{P}_n^{\,l} \le P(H_0|x) < \infty) = \alpha, \tag{4}$$

где $\hat{P}_n^{\,l} = G_*^{-1}(\alpha)$.

Для оценки результатов сравнений дикторов, близких к целевому, предлагается использовать нижнее значение ОДИ, а для дикторов, отличающихся от целевого — верхнее. Классический bootstrap-метод предполагает независимость элементов исходной выборки, что неверно для случая, когда сравниваются звуковые файлы одного и того же диктора. Для решения этой проблемы предлагается использовать subset bootstrap [23].

Экспериментальные результаты. На рис. 4 показана зависимость вероятности совпадения дикторов от расстояния между файлами (сплошная кривая) и кривая (пунктир) доверительных границ (ДГ), показывающая положение границ односторонних доверительных интервалов для α= 0,95. Такой уровень доверительности означает, что отображаемая кривой доверительных границ $P(H_0|x)$ имеет значения „не хуже“ показанных на графике, по крайней мере, для 95% дикторов в обучающей базе данных. „Хуже“ и „лучше“ для пользователей САИ означает, что для заданного α положительное решение САИ о совпадении дикторов понимается только при условии $\hat{P}_n^{\,l} > 0,5$, а отрицательное только при выполнении условия $\hat{P}_n^{\,u} < 0,5$. Кривая ДГ для заданного уровня доверительности дает значение вероятности совпадения дикторов, минимальное — при условии принятия решения об их совпадении и максимальное — при условии принятии решения об их различии. При использовании ДГ возникает принципиально новая область возможных решений, в которой с заданным уровнем доверительности нельзя принять ни положительного, ни отрицательного решения (LR=1). На рис. 4 это зона для значений $x \in (0,26; 0,28)$.

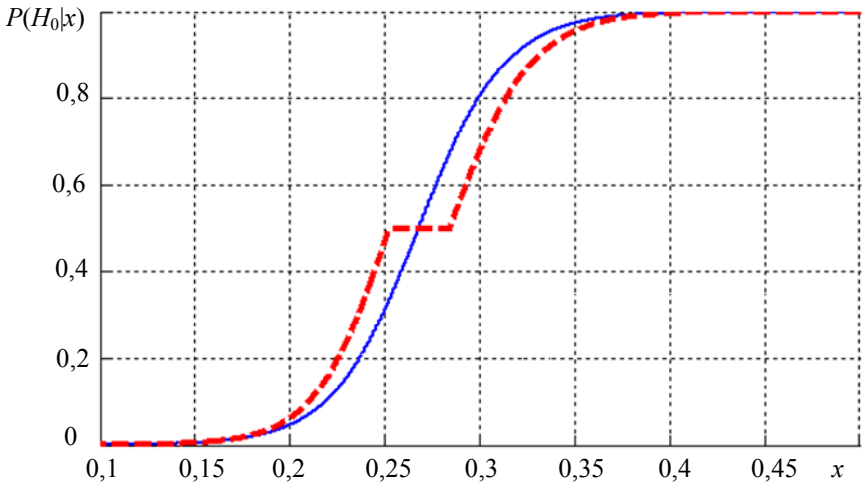
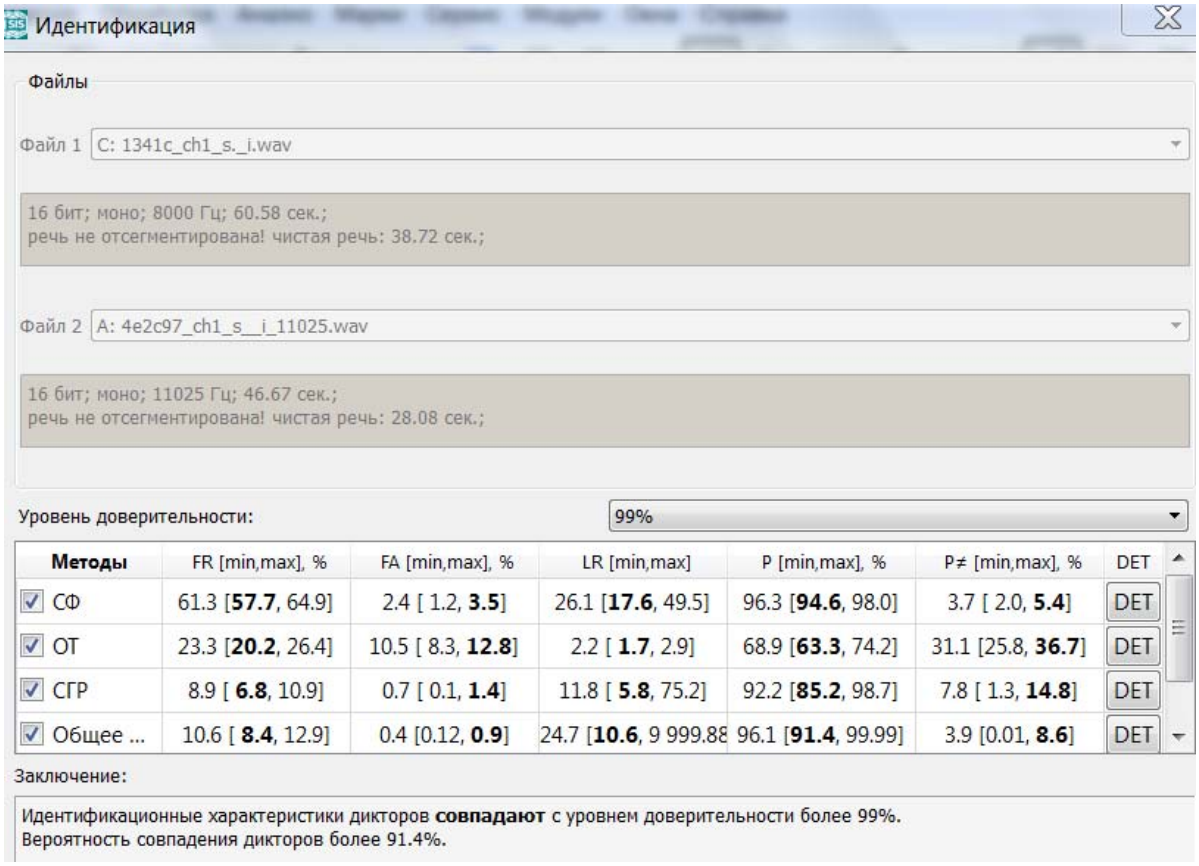


Рис. 4

Например, для данной САИ среднее значение апостериорной вероятности для значений $P(H_0|x_1)$ = 0,8 и $P(H_0|x_2)$ =0,98 соответствует значениям кривой ДГ $\hat{P}_n^{\,l}(H_0 \mid x_1) = 0,67$ и $\hat{P}_n^{\,l}(H_0 \mid x_2) = 0,97$. На рис. 5 приведены результаты идентификации системы SISII (STC 2012), вероятностные результаты сравнения дикторов для двух речевых файлов, ДИ и ДГ (жирный шрифт).

На рис. 6 приведены ДГ для двух разных баз: микрофонные интервью в NIST SRE 2008 [24] и телефонные диалоги в аналоговых ТСОП в базе RuSTeN [25]. $P(H_0|x)$ САИ [19] для баз данных NIST SRE 2008 (*1*) и RuSTeN (*2*) и кривые доверительных границ для $P(H_0|x)$ для NIST SRE 2008 (*3*) и RuSTeN (*4*), вычисленные при α= 0,95. Кривая ДГ показывает значения $P(H_0|x)$ для 95 % „лучших“, с точки зрения принимаемого идентификационного решения, дикторов в обучающей базе.



Методы	FR [min,max], %	FA [min,max], %	LR [min,max]	P [min,max], %	P≠ [min,max], %	DET
СФ	61.3 [**57.7**, 64.9]	2.4 [1.2, **3.5**]	26.1 [**17.6**, 49.5]	96.3 [**94.6**, 98.0]	3.7 [2.0, **5.4**]	DET
ОТ	23.3 [**20.2**, 26.4]	10.5 [8.3, **12.8**]	2.2 [**1.7**, 2.9]	68.9 [**63.3**, 74.2]	31.1 [25.8, **36.7**]	DET
СГР	8.9 [**6.8**, 10.9]	0.7 [0.1, **1.4**]	11.8 [**5.8**, 75.2]	92.2 [**85.2**, 98.7]	7.8 [1.3, **14.8**]	DET
Общее ...	10.6 [**8.4**, 12.9]	0.4 [0.12, **0.9**]	24.7 [**10.6**, 9 999.88	96.1 [**91.4**, 99.99]	3.9 [0.01, **8.6**]	DET

Рис. 5

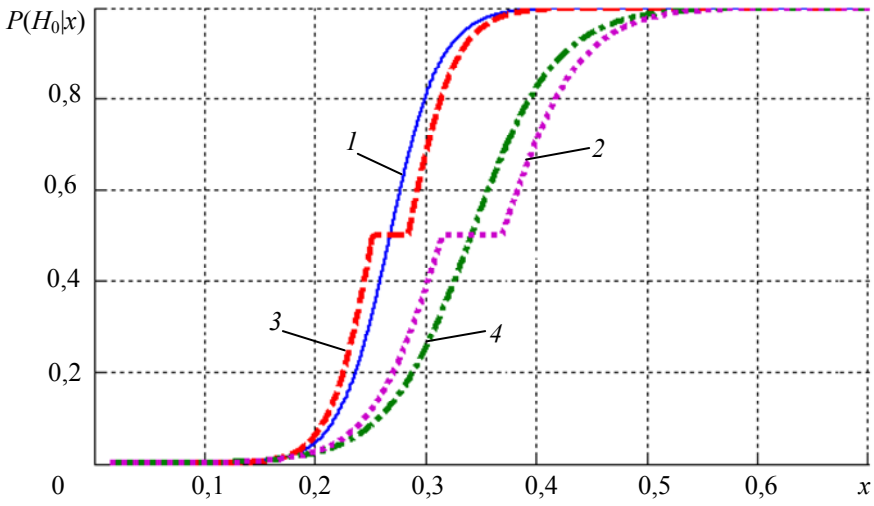


Рис. 6

Приведенные данные показывают высокую степень зависимости результатов работы САИ от типа обучающей базы данных, что требует наличия в пользовательском интерфейсе САИ настроек на свойства звуковых файлов.

Заключение. В работе предложены разработанные авторами, проверенные на практике подходы к организации пользовательского интерфейса САИ, ориентированные на решение задач мониторинга, учета и выполнения судебных экспертиз.

СПИСОК ЛИТЕРАТУРЫ

1. АПК „PhonoBase“ [Электронный ресурс]: <http://www.sis-tss.ru/2010-06-23-20-32-10/2326-qphonobaseq.html>.
2. [Электронный ресурс]: <http://www.agnitio.es/producto.php?id_producto=3>.
3. [Электронный ресурс]: <http://www.loquendo.com/en/products/speaker-verification/>.
4. *Попов Н. Ф.* и др. Идентификация лиц по фонограммам русской речи на автоматизированной системе „Диалект“. М., 1996.
5. *Тимофеев И. Н.* и др. Применение автоматизированной системы „Диалект“ на базе компьютерной речевой лаборатории CSL (США) при решении задач идентификации дикторов: Метод. рекомендации. ЭКЦ МВД РФ, 2000.
6. *Martin A. F., Greenberg C. S.* The NIST 2010 Speaker Recognition Evaluation // INTERSPEECH 2010. Makuhari, Chiba, Japan, 2010. P. 2726—2729.
7. *Wu J. C., Martin A. F., Kacker R. N.* Measures, Uncertainties, and Significance Test in Operational ROC Analysis // J. Res. NIST. 2011. Vol. 116, N 1. P. 517—537.
8. *Campbell W. M.* et al. Estimating and evaluating confidence for forensic speaker recognition // Proc. ICASSP2005. Philadelphia, PA, 2005.
9. *Rose P.* Technical forensic speaker recognition: Evaluation, types and testing of evidence // Computer Speech and Language. 2006. Vol. 20, N 2—3. P. 159—191.
10. *Belykh I. N.* et al. The speaker identification system for the NIST SRE 2010 // Informatics and its Applications. 2012. Vol. 6, N 1. P. 91—98.
11. *Drygajlo A.* Forensic automatic speaker recognition // IEEE Signal Processing Magazine. 2007. Vol. 24, N 2. P. 132—135.
12. *Drygajlo A.* Statistical Evaluation of Biometric Evidence in Forensic Automatic Speaker Recognition // IWCF 2009. Hague, Netherlands, 2009.
13. Interspeech 2008 special session “Forensic Speaker Recognition Traditional and Automatic Approaches“ [Электронный ресурс]: <http://interspeech2008.forensic-voice-comparison.net>.
14. *Зубова П. И., Коваль С. Л.* Методика экспертной идентификации дикторов по голосу и речи на основе комплексного анализа фонограмм // Теория и практика судебной экспертизы. 2007. Т. 3, № 7. С. 68—76.
15. *Evett I., Buckleton J.* Some aspects of the Bayesian approach for evidence evaluation // J. of Forensic Science Society. 1989. Vol. 29. P. 317—324.
16. *Meuwly D., Drygajlo A.* Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modeling // Proc. „Odyssey“. 2001. P. 145—150.
17. *Gonzalez-Rodriguez J.* et al. Robust likelihood ratio estimation in Bayesian forensic speaker recognition // Proc. Eurospeech. 2003. P. 693—696.
18. Guide to the Expression of Uncertainty in Measurement. Geneva, ISO, 1993.
19. *Koval S., Lokhanova A.* Confidence Bounds Curves as a Tool for Evaluation of Automatic Speaker Recognition Results Uncertainty // Proc. 14th Intern. Conf. on Speech and Computer. SPECOM 2011. Kazan, 2011. P. 284—289.
20. *Wu J., Martin A. F., Greenberg C. S., Kacker R. N.* Measurement Uncertainties in Speaker Recognition Evaluation // NIST Publication. 2010. P. 7722.
21. *Platt J.* Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods // Advances in Large Margin Classiers. Cambridge: MIT Press, 1999.
22. *Bolle R.M.* et al. Error Analysis of Pattern Recognition Systems: the Subsets Bootstrap // Computer Vision and Image Understanding. 2004. Vol. 93, N 1. P. 1—33.
23. *Efron B., Tibshirani R. J.* An Introduction to the Bootstrap. NY, 1993.
24. [Электронный ресурс]: < http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2011S11>.
25. LDC 2006S34 [Электронный ресурс]: <www.ldc.upenn.edu>.

Сведения об авторах

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

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

УДК 004.83

Ю. Н. МАТВЕЕВ

ОЦЕНКА ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА ОБЩЕГО РЕШЕНИЯ АНСАМБЛЯ КЛАССИФИКАТОРОВ

Предложен алгоритм оценки доверительного интервала общего решения ансамбля классификаторов, выходом каждого из которых является логарифмическое отношение правдоподобия.

Ключевые слова: *доверительный интервал, общее решение, ансамбль классификаторов, идентификация дикторов, голосовая биометрическая система.*

Введение. При построении голосовых биометрических систем для повышения надежности идентификации личности по голосу (идентификации диктора) часто используется набор автоматических, полуавтоматических и экспертных методов исследования фонограмм, основанных на признаках речи различной природы. В качестве надежного итогового решения используется обобщающее, построенное на базе решений каждого из перечисленных методов идентификации.

По результатам исследования фонограмм для каждого метода выдается мера доказательности, которая строится на основе оценки степени тождества/различия дикторов. В качестве меры доказательности принят логарифм отношения правдоподобия (*LR*-оценка, Likelihood Ratio) каждого метода:

$$\log(LR) = \log\left(\frac{P(X \mid H_0)}{P(X \mid H_1)}\right), \qquad (1)$$

где $P(X|H_0)$ — вероятность получения данных исследования X при истинности гипотезы H_0, $P(X|H_1)$ — вероятность получения данных исследования X при истинности гипотезы H_1.

Для автоматических методов идентификации оценка значений $P(X|H_0)$, $P(X|H_1)$ производится без участия эксперта. Для получения зависимости вероятностей $P(X|H_0)$, $P(X|H_1)$ от расстояния между векторами признаков сравниваемых сигналов на этапе разработки используются большие речевые базы данных, на которых устанавливается статистическая связь вероятностей с расстоянием. Тем не менее представительность этих баз данных ограничена условиями, в которых собирались фонограммы. В силу этого для условий, отличных от тех, на которые алгоритм был точно настроен, реальные значения вероятности истинности того или иного идентификационного решения становятся меньше. Наиболее важными факторами, влияющими на надежность решения автоматической системы, являются различия в свойствах канала записи эталонной (образцовой) и исследуемой (спорной) фонограмм, различное фи-

зиологическое и эмоциональное состояние диктора, несопоставимый речевой материал или условия внешней среды.

Для учета влияния этих факторов следует выявить их наличие и ввести к *LR*-оценке в формуле (1) степенную поправку, например, в форме степенного показателя *Q*:

$$LR = \left(\frac{P(X \mid H_0)}{P(X \mid H_1)} \right)^Q .$$

Эта степенная поправка учитывает сопоставимость исследуемых фонограмм с теми факторами, на которые ориентирован выбранный метод идентификации. Значение Q=1 соответствует тому, что качество фонограмм полностью удовлетворяет заявленным требованиям. Значение Q, близкое к нулю, соответствует тому, что качество фонограмм значительно ниже того, при котором метод сохраняет работоспособность; *LR*-оценка стремится к единице, а значит, невозможно принять решение по идентификации диктора.

Формула получения итогового обобщающего решения, т.е. итоговой оценки, наиболее часто реализуется в виде логарифмического отношения правдоподобия:

$$LLR = \log(LR) = \sum_i \log(LR_i), \qquad (2)$$

где $LR_i = FRR_i / FAR_i$ — оценка отношения правдоподобия (*LR*-оценка) *i*-го метода идентификации, FRR_i и FAR_i — оценки вероятностей ошибок первого и второго рода *i*-го метода идентификации соответственно.

Согласно руководству [1], результат измерения является только аппроксимацией или оценкой значения измеряемой величины и, таким образом, будет полным, только когда дополняется установлением неопределенности этой оценки. В соответствии с этим ставится задача установления неопределенности итоговой *LLR*-оценки.

На практике существует много возможных источников неопределенности, например, неполное определение и несовершенная реализация определения измеряемой величины; нерепрезентативная выборка измерений; неполное представление о влиянии условий окружающей среды на измерения или несовершенное измерение параметров окружающей среды; недостоверные значения констант и других параметров, полученных из внешних источников и используемых в алгоритме обработки данных; аппроксимации и предположения, используемые в методе измерения и измерительной процедуре и т.д. [2, раздел 3.3.2]. Практически все перечисленные источники неопределенности присутствуют при идентификации дикторов.

Согласно руководству [1], неопределенность — параметр, связанный с результатом измерения (оценкой), характеризующий дисперсию значений, которые могли быть обоснованно приписаны измеряемой величине. Параметром может быть, например, стандартное отклонение или полуширина интервала, имеющего установленный доверительный уровень. Стандартные отклонения оценивают из предполагаемых распределений вероятностей, основанных на опыте или другой информации.

Доверительное оценивание. В настоящее время в системах распознавания (верификации и идентификации) диктора все чаще применяется концепция доверительности [3—7]. В этом случае дополнительно определяется доверительная вероятность (доверительный интервал), указывающая на надежность полученного результата (оценки).

В соответствии с ГОСТ доверительный интервал с заданной вероятностью накрывает неизвестное значение оцениваемого параметра распределения. Границы доверительного интервала называют доверительными границами. Оцениванием с помощью доверительного интервала называют способ оценки, при котором с заданной доверительной вероятностью устанавливают границы доверительного интервала.

Несмотря на то что доверительные интервалы обычно применяют при оценке одного числового параметра, для многих двухпараметрических и трехпараметрических распределений (в задачах распознавания дикторов — нормальных, бинормальных, гамма-распределений) обычно используют точечные оценки и построенные на их основе доверительные границы для каждого из параметров отдельно.

Исходя из формулы (2) алгоритм оценки доверительного интервала итоговой *LLR*-оценки состоит из следующих шагов.

1) Оценка доверительных интервалов значений FAR_i и FRR_i для отдельных методов идентификации дикторов.

2) Оценка доверительных интервалов LR_i-оценок отдельных методов идентификации дикторов.

3) Оценка доверительного интервала итоговой *LLR*-оценки.

Основные подходы к оценке доверительных интервалов FAR и FRR. В литературе рассматривается несколько подходов к оценке значений *FAR* и *FRR*. В общем случае эти подходы делятся на параметрические и непараметрические методы оценки.

Непараметрические методы, наиболее популярными из которых являются методы бутстрепа, „блочного“ бутстрепа и т.д. [3, 8], требуют проведения множества тестов на различных выборках, что не подходит для экспертных и полуавтоматических методов идентификации дикторов из-за большой трудоемкости.

В параметрических методах используется предположение о виде распределения значений *FAR* и *FRR* при определении доверительных интервалов. Наиболее часто в работах по распознаванию дикторов делаются следующие предположения [9]:

— о бинормальном распределении,

— о нормальном распределении,

— о биномиальном распределении.

Допустим, что определены некоторое выборочное распределение D и некоторый доверительный порог δ. В предположении о бинормальном распределении генерирование доверительных интервалов и их границ производится с использованием доверительных границ Хотеллинга [3].

В предположении о нормальном распределении генерирование доверительных интервалов и границ производится путем расчета среднего μ и стандартного отклонения σ распределения D. Затем ищется статистическая константа z двусторонней доверительной границы δ для распределения размерности $|D|$, что дает доверительный интервал $\mu \pm z\sigma$.

В предположении о биномиальном распределении дисперсия рассчитывается как $V = \mu(1-\mu)$, что дает доверительный интервал $\mu \pm z\sqrt{V/|D|}$.

На практике наиболее часто используется предположение о биномиальном распределении. Решение, приведенное в работе [10], основывается на параметрической оценке доверительных интервалов значений *FRR* и *FAR* по методу, описанному в работе [11].

Оценка доверительных интервалов FAR и FRR оценок отдельных методов идентификации дикторов. Введем следующие обозначения: $P_к$ — априорная вероятность появления целевой личности (клиента), $P_з = 1 - P_к$ — априорная вероятность появления злоумышленника (импостера).

Значение этих параметров зависит от типа приложения. Например, в случае идентификации диктора при радиопередаче предполагается $P_к < 1$, в то время как в системах контроля доступа предполагается $P_з << 1$.

Определим доверительные интервалы для значений *FAR* и *FRR* в предположении о биномиальном характере распределения этих ошибок.

При большом числе попыток N распределение биномиальной случайной величины будет близко к нормальному. Учитывая, что дисперсия биномиальной случайной величины

равна $Np(1–p)$, получаем для ее математического ожидания Np приближенные доверительные границы для значения p (см. раздел 2.2.4 работы [12]):

$$\overline{p} - z\sqrt{\frac{\overline{p}(1-\overline{p})}{N}} \le p \le \overline{p} + z\sqrt{\frac{\overline{p}(1-\overline{p})}{N}}, \qquad (3)$$

где $\overline{p}$ — выборочная оценка p, а z — квантиль нормального распределения, равная хвосту кривой распределения $\alpha/2$ (например, $z = 1{,}96$ для $\alpha = 0{,}05$). Используя формулу (3), можно вычислить доверительный интервал частоты FAR (при $\overline{p} = FAR$ и $N=N_з$) и FRR (при $\overline{p} = FRR$ и $N=N_к$).

Оценка доверительных интервалов *LR*-оценок отдельных методов идентификации дикторов. В соответствии с формулой (2) следующим этапом является оценка доверительных интервалов LR_i-оценок отдельных методов идентификации дикторов, а точнее логарифма этих значений — LLR_i.

Оценка доверительного интервала значений LLR_i определяется по методу распространения ошибок [13]. Для случая функциональной зависимости вида $f = \frac{x}{y}$ $\left(LR_i = \frac{FRR_i}{FAR_i}\right)$ оценка доверительного интервала f вычисляется по следующей формуле:

$$\frac{\delta_f}{f} = \sqrt{\left(\frac{\delta_x}{x}\right)^2 + \left(\frac{\delta_y}{y}\right)^2}. \qquad (4)$$

Однако, поскольку при получении оценки LLR_i используется операция логарифмирования, то, как отмечено в работе [13], для этого должен использоваться другой подход к вычислению итоговой оценки, без вычисления выражения (4).

Вычисление доверительного интервала итоговой *LLR*-оценки. Для вычисления итоговой LLR оценки будем использовать абсолютные значения изменения величин FAR_i и FRR_i, т.е. ΔFAR_i и ΔFRR_i, соответственно. Предполагается, что величины FAR_i и FRR_i изменяются в диапазоне $(FRR_i – \Delta RRR_i\ ;\ FRR_i + \Delta FRR_i)$ и $(FAR_i – \Delta FAR_i\ ;\ FAR_i + \Delta FAR_i)$ соответственно.

Задача состоит в определении доверительного интервала величины LLR $(LLR–\Delta_d LLR;\ LLR + \Delta_u LLR)$, где $\Delta_d LLR$ и $\Delta_u LLR$ — абсолютное значение изменения величины LLR в сторону уменьшения и увеличения соответственно.

Воспользуемся формулой из работы [13]:

$$\Delta LLR = \sqrt{\sum_i \left(\Delta LLR^2_{FRR_i} + \Delta LLR^2_{FAR_i}\right)},$$

где ΔLLR_{FRRi} и ΔLLR_{FARi} — значение изменения LLR при изменении FRR_i и FAR_i, если остальные величины остаются неизменными.

Значение LLR определяет принадлежность тестового произнесения искомому диктору. При $LLR > 0$ тестовое произнесение принадлежит искомому диктору, $LLR < 0$ — нецелевому (злоумышленнику). Модуль значения LLR соответствует степени уверенности (большее значение по модулю соответствует большей уверенности).

Следовательно, при $LLR > 0$ наибольший интерес представляет величина $\Delta_d LLR$, которую можно найти по формуле:

$$\Delta_L LLR = \sqrt{\sum_i \left(\Delta_L LLR^2_{FRR_i} + \Delta_L LLR^2_{FAR_i}\right)},$$

где

$$\Delta_d LLR_{FRR_i} = \sum_j w_j \ln\left(\frac{FRR_j}{FAR_j}\right) - \sum_{j\neq i} w_j \ln\left(\frac{FRR_j}{FAR_j}\right) - w_i \ln\left(\frac{FRR_i - \Delta FRR_i}{FAR_i}\right) =$$

$$= w_i \ln\left(\frac{FRR_i}{FAR_i}\right) - w_i \ln\left(\frac{FRR_i - \Delta FRR_i}{FAR_i}\right) = w_i \ln\left(\frac{FRR_i}{FRR_i - \Delta FRR_i}\right),$$

$$\Delta_d LLR_{FAR_i} = w_i \ln\left(\frac{FRR_i}{FAR_i}\right) - w_i \ln\left(\frac{FRR_i}{FAR_i - \Delta FAR_i}\right) = w_i \ln\left(\frac{FAR_i - \Delta FAR_i}{FAR_i}\right),$$

$$w_i = D_i.$$

Аналогично при $LLR < 0$ наибольший интерес представляет величина $\Delta_u LLR$, которую можно найти по формуле:

$$\Delta_u LLR = \sqrt{\sum_i \left(\Delta_u LLR^2_{FRR_i} + \Delta_u LLR^2_{FAR_i}\right)},$$

где

$$\Delta_d LLR_{FRR_i} = w_i \ln\left(\frac{FRR_i + \Delta FAR_i}{FAR_i}\right) - w_i \ln\left(\frac{FRR_i}{FAR_i}\right) = w_i \ln\left(\frac{FRR_i + \Delta FRR_i}{FRR_i}\right),$$

$$\Delta_d LLR_{FAR_i} = w_i \ln\left(\frac{FRR_i}{FAR_i - \Delta FAR_i}\right) - w_i \ln\left(\frac{FRR_i}{FAR_i}\right) = w_i \ln\left(\frac{FAR_i}{FAR_i - \Delta FAR_i}\right),$$

$$w_i = D_i.$$

Заключение. В статье описан алгоритм оценки доверительного интервала для общего решения ансамбля из нескольких классификаторов (методов идентификации дикторов): автоматических, полуавтоматических и экспертных методов исследования фонограмм, основанных на признаках речи различной природы.

Описанный алгоритм оценки общего доверительного интервала основан на определении доверительных интервалов ошибок первого и второго рода (*FAR* и *FRR*) различных методов идентификации, составляющих ансамбль, в предположении о биномиальном характере распределения этих ошибок, а также оценке методом распространения ошибок доверительного интервала общего решения ансамбля по доверительным интервалам ошибок первого и второго рода (ΔFAR_i и ΔFRR_i) каждого из методов ансамбля.

СПИСОК ЛИТЕРАТУРЫ

1. Руководство по выражению неопределенности измерения / Пер. с англ., под науч. ред. проф. *В. А. Слаева.* ВНИИМ им. Д. И. Менделеева, 1999.
2. *Походун А. И.* Экспериментальные методы исследований. Погрешности и неопределенности измерений: Учеб. пособие. СПб: СПбГУ ИТМО, 2006. 112 с.
3. *Vogt R., Sridharan S., Mason M.* Making confident speaker verification decisions with minimal speech // Proc. of Interspeech. Brisbane, Australia, 2008. P. 1405—1408.
4. *Campbell W., Reynolds D., Campbell J., Brady K.* Estimating and evaluating confidence for forensic speaker recognition // Proc. of ICASSP. Philadelphia, PA, USA, 2005. Vol. 1. P. 717—720.
5. *Huggins J. G. M.* Confidence metrics for speaker identification // Proc. of ICSLP. Denver, Colorado, USA, 2002. P. 1381—1384.
6. *Richiardi J., Prodanov P., Drygajlo A.* Speaker verification with confidence and reliability measures // Proc. of ICASSP. Toulouse, France, 2006. Vol. 1. P. 641—644.
7. *Richiardi J., Drygajlo A., Prodanov P.* Confidence and reliability measures in speaker verification // J. of the Franklin Institute. 2006. Vol. 343, N 6. P. 574—595.

8. *Koval S., Lokhanova A.* Confidence Bounds Curves as a Tool for Evaluation of Automatic Speaker Recognition Results Uncertainty // Proc. 14th Intern. Conf. on Speech and Computer. SPECOM 2011. Kazan, 2011. P. 284—289.

9. *Wu J. C., Martin A. F., Kacker R. N.* Measures, Uncertainties, and Significance Test in Operational ROC Analysis // J. of Research of the National Institute of Standards and Technology. 2011. Vol. 116, N 1. P. 517—537.

10. Biosecure Tool. Performance evaluation of a biometric verification system, version 1.0. France, Aurelien Mayoue: GET-INT. 2007.

11. *Bolle R. M., Ratha N. K., Pankanti S.* Error analysis of pattern recognition systems — the subsets bootstrap // Computer Vision and Image Understanding. 2004. Vol. 93, N 1. P. 1—33.

12. *Мятлев В. Д., Панченко Л. А., Терехин А. Т.* Основы математической статистики. М.: МАКС Пресс, 2002.

13. Lab Reference Manual (LR09): Propagation of Uncertainty [Электронный ресурс]: <http://www.physics.pomona.edu/sixideas/labs/LRM/LR09.pdf>.

Сведения об авторе

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

Рекомендована кафедрой речевых информационных систем

Поступила в редакцию 22.10.12 г.

SUMMARY

P. 5—10.

A PHONETICALLY RICH TEXT FOR FUNDAMENTAL AND APPLIED RESEARCH ON RUSSIAN SPEECH VARIABILITY

A phonetically rich text intended to be used for research into regional and individual variability of Russian speech is presented. The text provides full coverage of basic phonetic units of Russian, which allows for application in fundamental and applied studies of various kind in the field of speech science.

Keywords: phonetically rich text, phonetically balanced text, statistical properties of russian speech, frequency and distribution of phonetic units.

Data on authors

Natalia S. Smirnova — Cand. Philolog. Sci.; STC Ltd., St. Petersburg; Head of the Linguistic Research Group; E-mail: nsmirnova@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. 11—18.

SUPPRESSION OF ACOUSTIC NOISE IN AUDIO DEVICE USING ASYNCHRONOUS REFERENCE SIGNAL

Semi-automatic technique for two-channel noise suppression with asynchronous reference signal (i.e. from an external source) is presented. The technique is described in details, its efficiency is compared with the algorithms using synchronous noise recordings.

Keywords: noise suppression, acoustic noise, adaptive signal processing.

Data on authors

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

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

P. 18—24.

ALGORITHMS FOR DETECTION OF TYPICAL NOISES AND INTERFERING BURSTS IN SPEECH SIGNALS

Methods of typical additive interfering noises and bursts detection in speech processing systems are analyzed and discussed. Detectors influence to the performance of speaker verification system is investigated experimentally. New improved algorithms for typical noises detection are proposed.

Keywords: noise, acoustic interference, pulse noise, speech signal processing.

Data on authors

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

Konstantin K. Simonchik — Cand. Techn. Sci.; STC Ltd., Department of Speaker Verification and Identification, St. Petersburg; Head of the Department; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; Associate Professor;
E-mail: simonchik@speechpro.com

P. 24—28.

MODERN MOBILE SYSTEM FOR TRACK WARNING

The track warning systems for railway workers travel teams performing tracks repair are analyzed. An alternative mobile warning system is presented. The system is based on remote sensors detecting train approach. A comparative analysis of the proposed version of the warning system and foreign systems is carried out.

Keywords: track warning system, train approach, vibroacoustic waves.

Data on authors

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

Maxim E. Markisonov — STC Ltd., St. Petersburg; Senior Sales Manager; E-mail: mme@speechpro.com

Sergey A. Panasyuk — Russian Railways Labor Protection and Industrial Safety Directorate, Moscow; Chief Specialist; E-mail: panasyuksa@gmail.com

P. 29—32.

AUTOMATION OF NEW VOICE CREATION PROCEDURE FOR A RUSSIAN TTS SYSTEM

An automatic system for creating a new voice for VitalVoice TTS is presented. The system includes text selection, speech recording and record monitoring, database labeling, and parameter setting of unit selection.

Keywords: speech synthesis, voice building, voice creation, automatic markup, diphone, text corpus.

Data on authors

Anna I. Solomennik — Speech Technology Ltd., Minsk; Scientist; E-mail: solomennik-a@speechpro.com

Pavel G. Chistikov — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems;
E-mail: chistikov@speechpro.com

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

Andrey O. Talanov — Cand. Techn. Sci.; STC Ltd., St. Petersburg, Speech Synthesis Department; Head of the Department; E-mail: andre@speechpro.com

Natalia A. Tomashenko — STC Ltd., St. Petersburg; Junior Scientist; E-mail: tomashenko-n@speechpro.com

P. 33—38.

A HYBRID TECHNOLOGY FOR TTS SYSTEM BASED ON HIDDEN MARKOV MODELS AND UNIT SELECTION ALGORITHM

An approach to synthesis of Russian TTS system based on integration of Hidden Markov Models and Unit Selection algorithms is presented. The voice model creation method is developed for constructing a natural intonation contour. Improved quality of synthesized speech is confirmed by experimental results.

Keywords: speech synthesis, Hidden Markov Models, Unit Selection, voice model.

Data on authors

Pavel G. Chistikov — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: chistikov@speechpro.com

Evgeny A. Korolkov — STC Ltd., St. Petersburg; Scientist; E-mail: korolkov@speechpro.com

Andrey O. Talanov — Cand. Techn. Sci.; STC Ltd., St. Petersburg, Speech Synthesis Department; Head of the Department; E-mail: andre@speechpro.com

Anna I. Solomennik — Speech Technology Ltd., Minsk; Scientist; E-mail: solomennik-a@speechpro.com

P. 38—42.

ASSESSMENT OF SYNTHESIZED SPEECH QUALITY: PROBLEMS AND SOLUTIONS

Various aspects of the speech synthesis systems quality assessment and comparison of existing TTS systems are concerned. A brief review of existing methods of quality assessment is presented.

Keywords: speech synthesis, quality of synthesized speech, TTS evaluation.

Data on authors

Anna I. Solomennik — Speech Technology Ltd., Minsk; Scientist; E-mail: solomennik-a@speechpro.com

Andrey O. Talanov — Cand. Techn. Sci.; STC Ltd., St. Petersburg, Speech Synthesis Department; Head of the Department; E-mail: andre@speechpro.com

Mikhail V. Solomennik — Cand. Techn. Sci.; Speech Technology Ltd., Minsk; Leading Software Engineer; E-mail: solomennik-m@speechpro.com

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

Pavel G. Chistikov — Post-Graduate Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: chistikov@speechpro.com

P. 42—46.

APPLICATION OF LINGUISTIC ANALYSIS FOR TEXT NORMALIZATION AND HOMONYMY RESOLUTION IN RUSSIAN TEXT-TO-SPEECH SYSTEM

A method based on automatic morphological and syntactic analysis is developed to resolve ambiguities that arise in the process of text normalization and homonymy resolution in VitalVoice Russian TTS system. A high degree of accuracy is demonstrated in experimental processing of Russian texts of various types.

Keywords: speech synthesis, linguistic analysis, text processing, text normalization, homonymy resolution.

Data on authors

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

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

Ilya M. Anichkin — STC Ltd., St. Petersburg; Leading Programmer; E-mail: anichkin@speechpro.com

P. 47—51.

STUDY OF INFORMATIVE SPEECH FEATURES FOR AUTOMATIC SPEAKER IDENTIFICATION

The most popular speech features used in automatic speaker recognition systems are studied. Results of experiments with speech database collected in different acoustic environments (wide range of signal/noise levels and reverberation times) and over different channels are reported.

Keywords: speech features, speaker identification.

Data on author

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

P. 51—61.

COMPARISON OF VARIOUS MIXTURES OF GAUSSIAN PLDA-MODELS IN THE PROBLEM OF TEXT-INDEPENDENT SPEAKER VERIFICATION

Applicability of unsupervised mixtures of PLDA models with Gaussian priors in a i-vector space for speaker verification is studied. Conditions under which the application is advantageous are analyzed for existing training databases. A mixture of two PLDA models is shown to be more effective than a single PLDA model for a cross-channel task.

Keywords: i-vector, joint factor analysis, PLDA mixture, speaker verification.

Data on authors

Timur S. Pekhovsky — Cand. Phys.-Math. Sci.; STC-Innovation Ltd., St. Petersburg; Leading Scientist; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; Associate Professor; E-mail: tim@speechpro.com

Alexander Yu. Sizov — Student; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; E-mail: sizov@speechpro.com

P. 61—66.

GINI CRITERION SVM FOR EMOTION CLASSIFICATION FRAMEWORK

Gini criterion is applied for creation of SVM classifier feature space. An experimental study of the optimal set of informative features and the classifier construction is presented.

Keywords: speech, emotion classification, Gini criterion, support vector machines.

Data on authors

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

Andrey G. Davydov — Cand. Techn. Sci.; Speech Technology Ltd., Minsk; Senior Scientist; E-mail: davydov-a@speechpro.com

Vitaliy V. Kiselyov — Speech Technology Ltd., Minsk; Director; E-mail: kiselev-v@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. 66—74.

FEATURES OF HUMAN-MACHINE INTERFACE OF MODERN BIOMETRIC IDENTIFICATION SYSTEMS

Modern systems designed for automated identification of personality based on biometric characteristics analysis is considered. Requirements on arrangement of human-machine interface for such systems are formulated.

Keywords: human-machine interface, person identification, biometric identification system, voice biometry.

Data on authors

Dmitry V. Dyrmovsky — STC Ltd., Moscow; Director of the Branch; St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Department of Speech Information Systems; Post-Graduate Student; E-mail: ddv@speechpro.com

Sergey L. Koval — Cand. Techn. Sci.; STC Ltd., Moscow; Principle Expert; E-mail: koval@speechpro.com

P. 74—79.

EVALUATION OF THE CONFIDENCE INTERVAL FOR DECISION PREDICTION OF AN ENSEMBLE OF CLASSIFIERS

An algorithm is proposed for evaluation of the confidence interval for decision prediction of an ensemble of classifiers, where each classifier in the ensemble returns a prediction as a logarithmic likelihood ratio.

Keywords: confidence interval, decision prediction, ensemble of classifiers, speaker identification, voice biometric system.

Data on author

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