УНИВЕРСИТЕТ ИТМО
НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2015 Том 15 № 1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2015 Vol. 15 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 621.391.037.372

МЕТОД ИДЕНТИФИКАЦИИ ДИКТОРОВ НА ОСНОВЕ СРАВНЕНИЯ СТАТИСТИК ДЛИТЕЛЬНОСТЕЙ ФОНЕМ

Е.В. Булгакова[a], А.В. Шолохов[b], Н.А. Томашенко[a, c]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] Университет Восточной Финляндии, Йоэнсуу, FI-80101, Финляндия
[c] ООО «Центр речевых технологий», Санкт-Петербург, 196084, Российская Федерация
Адрес для переписки: bulgakova@speechpro.com

Информация о статье
Поступила в редакцию 27.11.14, принята к печати 05.12.14
doi: 10.17586/2226-1494-2015-15-1-70-77
Язык статьи – русский
Ссылка для цитирования: Булгакова Е.В., Шолохов А.В., Томашенко Н.А. Метод идентификации дикторов на основе сравнения статистик длительностей фонем // Научно-технический вестник информационных технологий, механики и оптики. 2015. Том 15. № 1. С. 70–77

Аннотация.
Предмет исследования. Представлен полуавтоматический метод идентификации диктора по речи на основе сравнения просодических признаков – статистик длительностей звуков. В последнее время благодаря развитию речевых технологий наблюдается значительный интерес к поиску экспертных методов идентификации диктора по голосу, дополняющих с целью повышения надежности идентификации известные методы, а также обладающих низкой трудоемкостью. Эффективное решение данной проблемы необходимо для принятия надежного решения о тождестве либо различии голосов дикторов, представленных на фонограммах.
Описание метода. Впервые представлен алгоритм расчета оценки различия голосов дикторов на основе сравнения статистик длительностей фонем и аллофонов. Характерной особенностью предложенного метода является возможность его применения в комплексе с другими полуавтоматическими методами (акустическими, аудитивно-лингвистическими) в связи с отсутствием ярко выраженной корреляции между анализируемыми признаками. Преимуществом метода является возможность проведения экспресс-исследования фонограмм большой длительности за счет автоматизации процесса подготовки данных для анализа. Описываются принципы работы автоматического сегментатора речи, используемого для расчета статистик длительностей звуков по акустико-фонетической разметке. Программное обеспечение разработано в качестве инструмента подготовки данных для экспертного анализа.
Апробация метода. Метод апробирован на базе 130 речевых записей, включающей русскую речь дикторов-мужчин и дикторов-женщин, и показал надежность 71,7% на базе, содержащей записи женской речи, и 78,4% на базе, содержащей записи мужской речи. Также было экспериментально установлено, что из всех используемых признаков наиболее информативными являются статистики длительностей фонем гласных и сонорных согласных.
Практическая значимость. Результаты эксперимента показали применимость предложенного метода для решения задачи распознавания диктора по голосу и речи в рамках проведения фоноскопической экспертизы.
Ключевые слова: фоноскопическая экспертиза, распознавание диктора, полуавтоматические методы идентификации, статистика длительностей фонем, фонемная сегментация.
Благодарности. Работа выполнена при государственной финансовой поддержке ведущих университетов Российской Федерации (субсидия 074-U01).

SPEAKERS' IDENTIFICATION METHOD BASED ON COMPARISON OF PHONEME LENGTHS STATISTICS

E.V. Bulgakova[a], A.V. Sholokhov[b], N.A. Tomashenko[a, c]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] University of Eastern Finland, Joensuu, FI-80101, Finland
[c] "Speech Technology Center", LLC, Saint Petersburg, 196084, Russian Federation
Corresponding author: bulgakova@speechpro.com

Article info
Received 27.11.14, accepted 05.12.14
doi: 10.17586/2226-1494-2015-15-1-70-77
Article in Russian
Reference for citation: Bulgakova E.V., Sholokhov A.V., Tomashenko N.A. Speakers' identification method based on comparison of phoneme lengths statistics. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2015, vol. 15, no. 1, pp. 70–77 (in Russian)

Abstract.
Subject of research. The paper presents a semi-automatic method of speaker identification based on prosodic features comparison - statistics of phone lengths. Due to the development of speech technologies in recent times, there is an increased interest in searching of expert methods for speaker's voice identification, which supplement existing methods to increase identification reliability and also have low labour intensity. An efficient solution for this problem is necessary for making the reliable decision whether the voices of the speakers in the audio recordings are identical or different.
Method description. We present a novel algorithm for calculating the difference of speakers' voices based on comparing of statistics for phone and allophone lengths. Characteristic feature of the proposed method is the possibility of its application along with the other semi-automatic methods (acoustic, auditive and linguistic) due to the lack of a strong correlation between analyzed features. The advantage of the method is the possibility to carry out rapid analysis of long-duration recordings because of preprocessing automation for data being analyzed. We describe the operation principles of an automatic speech segmentation module used for statistics calculation of sound lengths by acoustic-phonetic labeling. The software has been developed as an instrument of speech data preprocessing for expert analysis.
Method approbation. This method was approved on the speech database of 130 speech records, including the Russian speech of the male speakers and female speakers, and showed reliability equal to 71.7% on the database containing female speech records, and 78.4% on the database containing male speech records. Also it was experimentally established that the most informative of all used features are statistics of phone lengths of vowels and sonorant sounds.
Practical relevance. Experimental results have shown applicability of the proposed method for the speaker recognition task in the course of phonoscopic examination.
Keywords: phonoscopic examination, speaker recognition, semi-automatic speaker identification methods, statistics of phone lengths, phone segmentation.
Acknowledgements. The work is partially financially supported by the Government of the Russian Federation (grant 074-U01).

Введение

Информация, содержащаяся в речевом сигнале, позволяет решать одну из наиболее актуальных проблем современных речевых технологий – задачу распознавания диктора. Данная задача заключается в установлении тождества личности по совокупности общих и частных признаков его голоса и речи [1] и объединяет подзадачи идентификации и верификации дикторов. В первом случае спорная фонограмма сравнивается с набором образцов голоса для составления списка наиболее похожих голосов. Во втором случае спорная фонограмма сравнивается с образцом голоса с целью подтверждения идентичности голосов дикторов на представленных фонограммах. Представленный в настоящей работе метод может быть применим для обеих подзадач.

В настоящее время для решения задачи распознавания диктора широко применяются как автоматические (объективные), так и экспертные (субъективные) методы. Использование экспертных методов в процессе проведения фоноскопических исследований с целью идентификации говорящего дает возможность уточнить и скорректировать работу автоматических средств анализа и сравнения речевых сигналов. Кроме того, экспертные методы идентификации диктора применяются в случаях, когда работа автоматических методов оказывается затрудненной – например, в условиях сильной зашумленности фонограмм. Среди применяемых экспертных методов можно выделить акустические, лингвистические и аудитивные [2–8]. Данные методы предъявляют высокие требования к уровню квалификации эксперта и обладают значительной трудоемкостью. Предложенный полуавтоматический метод идентификации дикторов, основанный на сравнении статистик длительностей аллофонов фонем, позволяет эксперту работать с большим объемом данных в условиях проведения экспресс-анализа фонограмм.

Полуавтоматические методы идентификации дикторов по голосу

Подавляющее большинство современных методов идентификации дикторов по голосу и речи основаны на статистическом анализе распределения аудитивно-лингвистических или акустических признаков. Анализ современной методической и научной литературы по проблеме идентификации дикторов показал, что в современной экспертной практике применяются различные полуавтоматические акустические и аудитивно-лингвистические методы. Метод анализа мелодического контура, метод формантного выравнивания и микроанализ спектров гласных можно условно назвать акустическими.

Метод мелодического контура [5, 6] позволяет эксперту анализировать и сравнивать основные характеристики мелодических структур, представленные в виде наборов значений параметров основного тона для сопоставимых участков мелодического контура (опорных фрагментов). Возможность сравнения мелодического оформления различных фрагментов речевого сигнала обеспечивается их относительной реализационной стабильностью в сопоставимых контекстах, т.е. типичностью и повторяемостью в речи конкретного диктора с поправкой на характерную для него специфику контекстной и иной внутридикторской вариативности. Реализованный в настоящее время метод обладает точностью идентификации около 70% [6], но требует настройки под конкретный язык.

Метод формантного выравнивания [3, 4] и микроанализ спектров гласных [8, 9] являются модификацией распространенного за рубежом метода «voiceprint», который подвергся серьезной критике в кругу ученых по причине низкой точности идентификации [10]. В качестве недостатка данного метода можно

отметить высокую трудоемкость. Метод микроанализа спектров гласных носит текстозависимый и языкозависимый характер.

К аудитивно-лингвистическим методам относятся лингвистический [7] и аудитивный методы [2] идентификации дикторов. Реализованный в настоящее время лингвистический метод состоит из выявления и сравнения признаков как сегментного уровня, касающихся специфики произношения отдельных звуков (гласных и согласных) и их сочетаний в сопоставимых фонетических контекстах, так и признаков супрасегментного уровня, связанных с особенностями проявления фразовой интонации, ударения конкретного языка. Хотя данный вид анализа можно назвать универсальным, набор гласных и согласных фонем, типы фонетических процессов, так же как и мелодические особенности, сильно зависят от языка. Это значит, что лингвистический анализ речи в случае проведения фоноскопических исследований на незнакомом языке невозможен без ознакомления с данным языком даже для квалифицированного эксперта.

Аудитивный метод идентификации диктора [2] представляет собой способ формализации слухового впечатления от голоса и речи на основании сравнения фиксированного набора слуховых характеристик. Данный метод заключается в прослушивании подготовленным экспертом звукозаписей образца речи идентифицируемого диктора и спорной фонограммы, анализе звукового материала и выделении индивидуализирующих дикторских характеристик, подтверждающих или опровергающих факт принадлежности речи, записанной на образцовой и спорной фонограммах, одному и тому же диктору.

Проведение полного аудитивного анализа предполагает исследование полного набора аудитивных признаков [8], что требует значительных трудозатрат эксперта. Данный факт можно отнести к недостаткам метода.

Таким образом, каждый из перечисленных выше методов имеет ряд существенных недостатков:

- метод сравнения мелодического контура является языкозависимым и предъявляет высокие требования к сопоставимости материала по эмоциональному состоянию диктора;
- применение лингвистического метода затруднено в условиях анализа иноязычной речи диктора, записанного на фонограмме;
- аудитивный метод и метод формантного выравнивания требуют значительных трудозатрат эксперта (2–4 ч для обработки двух фонограмм длительностью 5 мин каждая);
- метод микроанализа спектров гласных является языкозависимым и текстозависимым.

Предлагаемый в настоящей работе альтернативный метод идентификации диктора, основанный на сравнении просодических признаков – статистик длительностей аллофонов фонем, дополняет работу описанных выше методов благодаря анализу признаков, не обладающих ярко выраженной корреляцией с признаками, применяемыми в других методах. Преимуществом данного метода является возможность его использования для экспресс-анализа фонограмм большой длительности. Время исследования двух фонограмм длительностью 30 мин каждая составляет от нескольких секунд до 10 мин (с учетом проверки и корректировки экспертом границ аллофонов фонем). К числу возможных достоинств этого метода предположительно относится его применимость в условиях анализа фонограмм, содержащих речь диктора на иностранных языках, в связи с тем, что существует некоторый универсальный набор фонем, присутствующий в различных языках мира [11]. Статистики длительностей данных фонем также могут быть использованы для идентификационного исследования фонограмм. Данная гипотеза, выдвинутая на основе предварительных результатов экспериментов, будет проверяться в последующих исследованиях.

Описание принципов работы метода

На первой стадии эксперимента с помощью автоматического сегментатора речи на основе звуковых файлов и заранее подготовленных текстовых расшифровок, соответствующих этим файлам, речевой массив данных был отсегментирован на фонемы. Далее с помощью специально созданного программного обеспечения был осуществлен расчет статистик длительностей аллофонов фонем. Мы рассматриваем аллофон как реализованный в речи вариант фонемы, обусловленный конкретным фонетическим окружением. На заключительной стадии эксперимента проводилось сравнение статистик идентичных аллофонов, присутствующих в речи дикторов. Основные этапы работы алгоритма представлены на рис. 1.

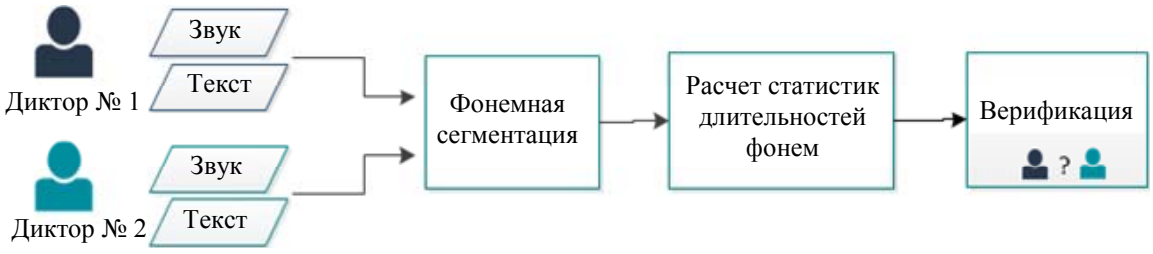

Рис. 1. Этапы работы алгоритма верификации на основе сравнения статистик длительности фонем

1. **Фонемная сегментация фонограмм, содержащих речь диктора № 1 и диктора № 2.** В процессе сегментации определяются временные границы каждого аллофона. После проведения автоматической сегментации в случае необходимости эксперт может откорректировать границы выделенных аллофонов.
2. **Вычисление средних длительностей для каждого аллофона по полученной сегментации.** При этом предполагается, что на фонограммах присутствует полный набор рассматриваемого множества аллофонов. В противном случае для недостающих аллофонов используются средние значения их длительностей, рассчитанные на дополнительной базе записей речи, принадлежащих многим различным дикторам, так называемой базе развития.
3. **Вычисление параметрической оценки различия голосов дикторов и принятие решения.** Параметры этой оценки вычисляются по базе развития. Кроме этого, оценивается пороговое значение оценки различия, которое используется для принятия решения о тождестве или различии голосов двух дикторов (верификация), присутствующих на паре фонограмм.

Описание фонемного сегментатора

Признаки для идентификации были получены с помощью фонемного автосегментатора [12]. При сегментации использовалась акустическая модель (АМ), параметры которой оценивались на базе объемом около 150 ч звучащей речи, состоящей из новостных передач, чтения и спонтанной речи на русском языке. АМ для сегментации представляет собой статистическую модель, построенную на основе скрытых марковских моделей (Hidden Markov Models, HMM) [13], в которых состояния моделей фонем или контекстных трифонов (фонем с определенным левым и правым контекстом) моделируются с помощью смеси гауссовых распределений (Gaussian Mixture Models, GMM) [13]. Каждое состояние трифона содержит три состояния. По типу извлекаемых акустических признаков данная АМ относится к классу тандемных акустических моделей, что означает, что в качестве акустических признаков используются контекстные признаки (Left Context – Right Context, LC-RC) [14]. Используется базовое множество из 54 фонетических классов, соответствующих фонемам русского языка и паузе. Из этого множества с применением процедуры связывания трифонов по фонетическому дереву находятся связанные состояния трифонов. Построенная таким образом АМ содержит всего 13700 связанных состояний. Каждое состояние трифона в АМ моделируется с помощью смеси гауссовых распределений, содержащей в среднем 14 гауссоид – это оптимальное количество для данной выборки, найденное с помощью алгоритма кроссвалидационного контроля состояний, описанного в работах [15, 16].

Фонемная сегментация выполняется автоматически с помощью модулей системы распознавания речи (Automatic Speech Recognition, ASR) с АМ, параметры которой описаны выше. Сегментация проводится на основе выравнивания (force alignment) транскрипции и звукового сигнала. Здесь под выравниванием подразумевается нахождение в звуковом файле моментов времени, соответствующих началам и концам произнесений слов и фонем.

Качество сигнала исследуемых фонограмм может быть недостаточно высоким для надежной работы классического алгоритма выравнивания Витерби [13], например, содержать шумы, неречевые фрагменты. Также сами фонограммы могут иметь большую длительность, а текстовки к ним – содержать неточности и ошибки. Все эти факторы могут приводить к сбою работы алгоритма Витерби. В связи с этим в описанной системе используется алгоритм выравнивания на основе метода опорных фрагментов [17]. Основная идея метода заключается в нахождении так называемых «островков надежности» – участков фонограммы, которые задают разбиение исходной фонограммы на более короткие опорные фрагменты, для каждого из которого известна текстовка и транскрипция. Поиск «островков надежности» осуществляется путем запуска системы ASR на имеющейся фонограмме и последующего сравнения результата распознавания с имеющейся текстовой расшифровкой. Сравнение текстов проводится методом динамического программирования на основе расстояния Левенштейна [18]. Далее на этих более коротких фрагментах можно либо запускать выравнивание по алгоритму Витерби, либо повторять процедуру рекурсивно, как это делается в [17]. В данной схеме также целесообразно использовать алгоритм, описанный в [12].

Расчет оценки различия голосов дикторов

Для принятия решения о тождестве или различии голосов дикторов на двух фонограммах необходимо ввести некоторую скалярную величину – оценку различия. Она будет определять, насколько близки речевые характеристики, извлеченные из пары записей речи.

Предположим, что $\mathbf{x}$ и $\mathbf{y}$ – вектора, содержащие средние длительности для каждого аллофона. Размерность этих векторов тогда будет равна общему количеству рассматриваемых аллофонов. Далее определим оценку различия для двух таких векторов, которую здесь и далее будем называть евклидовой:

$$s(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d} w_i (x_i - y_i)^2}\ , \qquad (1)$$

где w_i – весовые коэффициенты; d – общее число аллофонов. Из формулы (1) видно, что для частного случая $w_i = 1$ для всех i, эта оценка превращается в евклидову метрику. В общем же случае мы имеем взвешенную евклидову метрику $s(\mathbf{x},\mathbf{y}) = \sqrt{x^T \mathbf{W} y^T}$, которая определяется диагональной матрицей неотрицательных весов $\mathbf{W} = \mathrm{diag}[w_1, w_2, ..., w_d]$. Таким образом, чем меньше значение $s(\mathbf{x},\mathbf{y})$, тем более вероятно, что личности дикторов, говорящих на фонограммах, представленных векторами $\mathbf{x}$ и $\mathbf{y}$, совпадают.

Экспериментальные результаты

Эксперименты проводились на базе речевых записей, включающей русскую речь 26 дикторов (6 мужчин и 20 женщин), каждый из которых был записан в 5 подходах. Иными словами, общее количество записей мужской речи было равно 30, женской – 100.

Предложенный алгоритм идентификации диктора был реализован на языке Python. Для оценки эффективности метода идентификации диктора использовалось значение равенства ошибок I и II рода – равновероятная ошибка (Equal Error Rate, EER, %) [1]. Для вычисления EER было сформировано два типа сравнений: свой-свой и свой-чужой – отдельно для каждого пола дикторов.

Нами было рассмотрено два способа назначения весовых коэффициентов в оценке различия. В первом случае все веса были равны 1. Другими словами, ни один из признаков не имел преимущества в значимости перед остальными. Во втором случае веса вычислялись по формуле:

$$w_i = \left(\frac{D_i^{between}}{D_i^{total}} \right)^{\alpha}, \qquad (2)$$

где $D_i^{between}$ – межклассовая дисперсия i-ой координаты вектора признаков; D_i^{total} – дисперсия i-ой координаты вектора признаков; α – эмпирический степенной коэффициент, значение которого подбирается экспериментально. $D_i^{between}$ – дисперсия средних значений векторов-признаков, рассчитанных отдельно для каждого класса (под классом понимается совокупность всех векторов, соответствующих речи одного диктора).

Выбор такого вида весовых коэффициентов можно обосновать тем, что признаки с большей межклассовой дисперсией обладают большей дискриминирующей способностью. Иначе говоря, значение коэффициента должно быть пропорционально значению $D_i^{between}$. Однако в этом случае признаки с большой дисперсией могут стать более значимыми, вне зависимости от их информативности для классификации. В связи с этим значение межклассовой дисперсии нормируется на значение общей дисперсии D_i^{total}.

По формуле (2) были найдены 10 аллофонов, имеющих наибольший вес (от меньшего к большему), – [$у_{1\text{-й п/уд}}$], [н], [п'], [м], [н'], [$а_{уд}$], [л'], [$у_{уд}$], [$и_{з/уд}$], [$а_{1\text{-й п/уд}}$], а также 10 аллофонов, имеющих наименьший вес (от меньшего к большему), – [с], [$э_{уд}$], [т], [ш], [х'], [б'], [т'], [г'], [к'], [к], где апостроф обозначает мягкость согласного, а сокращения уд, п/уд, з/уд используются для обозначения положения гласного в зависимости от места ударения. Например, [$а_{уд}$] – гласный /а/ в ударной позиции, [$у_{1\text{-й п/уд}}$] – гласный /у/ в первом предударном слоге, [$и_{з/уд}$] – гласный /и/ в заударной позиции.

Приведенные результаты, несмотря на малый размер выборки, используемой для проведения исследования, позволяют сделать некоторые выводы:

1. большинство найденных аллофонов, имеющих наибольший вес, представляют собой гласные либо сонорные согласные (сонанты);
2. к числу аллофонов, имеющих наименьший вес, принадлежат глухие согласные.

Это свидетельствует о том, что наиболее информативными с точки зрения идентификации являются речевые участки, соответствующие гласным либо наиболее близким к ним по акустическим характеристикам сонорным согласным.

В таблице приведены значения EER для двух вариантов выбора оценки: евклидова метрика – равные веса, взвешенная евклидова метрика – веса, полученные из формулы (2). Как видно из таблицы, использование весов (2) повышает точность метода от 2,6% до 4%.

Пол	Оценка различия	EER, %
Мужчины	Евклидова метрика	25,56
	Взвешенная евклидова метрика	21,66
Женщины	Евклидова метрика	30,84
	Взвешенная евклидова метрика	28,29

Таблица. Показатели EER-идентификации по речи для двух полов дикторов

Кроме того, по полученным оценкам были построены ROC-кривые (Receiver Operating Characteristic) [19] (рис. 2), которые показывают зависимости доли верных положительных ответов от

общего числа положительных ответов True Positive Rate (TPR) и доли ошибочных положительных ответов (тождество дикторов) от общего числа отрицательных ответов (различие дикторов) False Positive Rate (FPR). Также была исследована зависимость эффективности системы от значения степенного коэффициента. На рис. 3 приведены графики, показывающие наличие явного минимума (в области значений α, равных 5–6).

Обсуждение результатов и дальнейшие перспективы

Как показывают результаты проведенного эксперимента, разработанный метод статистик длительностей аллофонов фонем может быть применим для решения задачи идентификации диктора по речи в связи с тем, что ошибка идентификации данного метода сопоставима с уровнем EER других полуавтоматических методов [6]. Метод был апробирован на записях, содержащих русскоязычный материал. Также был предложен способ расчета весовых коэффициентов, который позволяет улучшить производительность метода.

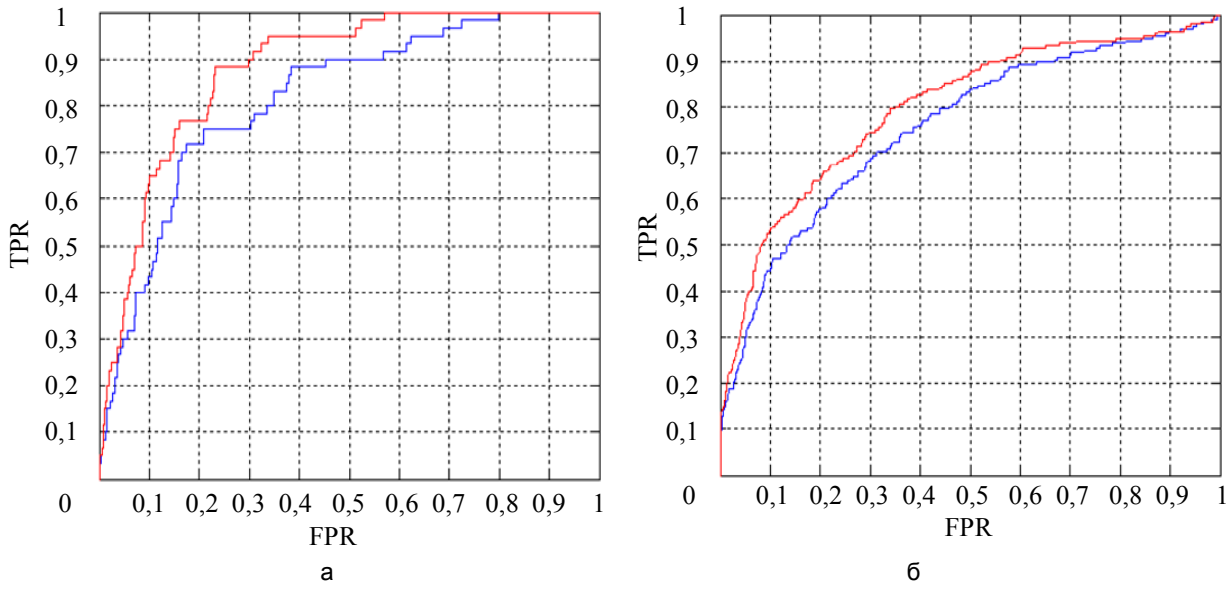

Рис. 2. ROC-кривые для мужчин (а) и женщин (б). Синий – евклидова метрика, красный – взвешенная евклидова метрика

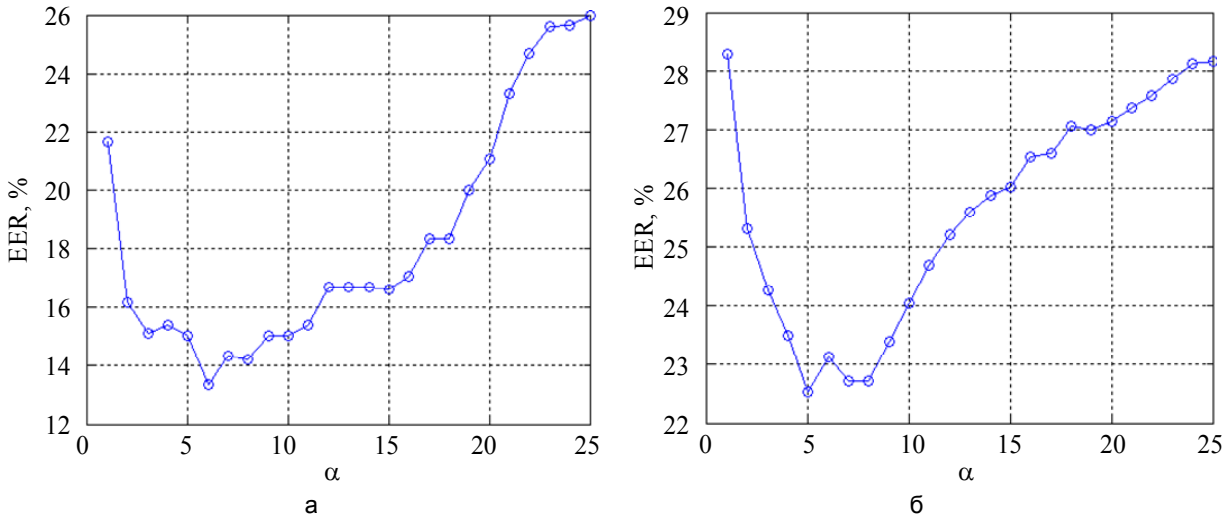

Рис. 3. Зависимость EER от значения коэффициента α для: мужчин (а); женщин (б)

Полученные данные подтверждают экспериментально установленный вывод [1] о том, что при верификации женских голосов алгоритмы, как правило, показывают худшую эффективность, чем в случае с мужскими голосами.

Следует отметить, что в представленном способе вычисления оценки никак не принимаются во внимание статистические зависимости между признаками – длинами аллофонов. В дальнейшем планируется расширить данную модель таким образом, чтобы она учитывала эти зависимости.

Другим недостатком описанной модели является невозможность работы в условиях пропущенных данных, т.е. когда один из аллофонов ни разу не встретился в записанном произнесении. В связи с этим планируется разработка вероятностной формулировки этой модели, которая позволит обрабатывать входные данные с пропусками. В дальнейшем планируется исследование эффективности работы данного метода с использованием больших объемов данных, при анализе фонограмм, содержащих иноязычную речь, а также в условиях повышенной зашумленности.

Заключение

В работе впервые предложен метод идентификации на основе сравнения статистик длительностей фонем, характерных для речи конкретного диктора. Рассмотрены особенности метода среди остальных полуавтоматических методов, проведены эксперименты, сделаны выводы и обозначены перспективы применения метода. Преимуществом метода является возможность его использования для проведения экспресс-анализа фонограмм большой длительности. Надежность метода составляет 71,7% на базе, содержащей записи женской речи, и 78,4% на базе, содержащей записи мужской речи. Также были выявлены наиболее информативные признаки по результатам ранжирования статистик длительностей русских фонем.

References

1. Kozlov A., Kudashev O., Matveev Y., Pekhovsky T., Simonchik K., Shulipa A. SVID speaker recognition system for the NIST SRE 2012. *Lecture Notes in Computer Science*, 2013, vol. 8113 LNAI, pp. 278–285. doi: 10.1007/978-3-319-01931-4_37
2. Prodan A.I., Talanov A.O. *Ispol'zovanie nabora slukhovykh kharakteristik rechi pri identifikatsii po golosu* [Using a hearing aid in the identification of the characteristics of speech voice]. *Materialy 14 Mezhdunarodnoi Konferentsii Speech and Computer, SPECOM'2011*. [Proc. 14th Int. Conf. Speech and Computer, SPECOM'2011]. Kazan', Russia, 2011, pp. 338–344.
3. Koval' S.L., Khitrov M.V. *Identifikatsiya diktorov pri analize raznoyazychnykh fonogramm na osnove sravneniya formantnykh spektrov* [Speaker identification in the analysis of multilingual tracks basedv on formant spectra comparison]. Available at: http://zhenilo.narod.ru/new_main/ips/2003_speech.pdf, свободный. Яз. рус. (accessed 7.11.2014).
4. Koval S. Formants matching as a robust method for forensic speaker identification. *Proc. 11th Int. Conf. on Speech and Computer*. St. Petersburg, 2006, pp. 125–128.
5. Smirnova N., Starshinov A., Oparin I., Goloshchapova T. Using parameters of identical pitch contour elements for speaker discrimination. *Proc. 12th Int. Conf. on Speech and Computer, SPECOM 2007*. Moscow, Russia, 2007, pp. 361–366.
6. Smirnova N.S. *Speaker identification based on the comparison of utterance pitch contour parameters*. Available at: http://www.dialog-21.ru/digests/dialog2007/materials/html/77.htm (accessed 7.11.2014) [In Russian].
7. Koval' S.L., Labutin P.V., Pekhovskii T.S., Proshchina E.A., Smirnova E.A., Talanov A.O. *Metodika identifikatsii diktorov po golosu i rechi na osnove kompleksnogo analiza fonogramm* [Technique of speaker identification by voice and speech based on a comprehensive analysis of phonograms]. Available at: http://www.dialog-21.ru/digests/dialog2007/materials/html/39.htm (accessed 7.11.2014)
8. Popov N.F., Lin'kov A.N., Kurachenkova N.B., Baicharov N.V. *Identifikatsiya Lits po Fonogrammam Russkoi Rechi na Avtomatizirovannoi Sisteme "Dialekt"* [Identification of Persons by Russian Speech Phonograms on the Automated System "Dialect"]. Moscow, Voiskovaya chast' 34435 Publ., 1996, 102 p.
9. Rose P. Speaker verification under realistic forensic conditions. *Proc. 6th Australian Int. Conf. on Speech Science and Technology*. Adelaide, South Australia, 1996, pp. 109–114.
10. Hollien H. *Forensic Voice Identification*. NY, Academic Press, 2001, 240 p.
11. Ladefoged P. *Preliminaries to Linguistic Phonetics*. Chicago, University of Chicago Press, 1971, 122 p.
12. Tomashenko N., Khokhlov Y. Fast algorithm for automatic alignment of speech and imperfect text data. *Lecture Notes in Computer Science*, 2013, vol. 8113 LNAI, pp. 146–153. doi: 10.1007/978-3-319-01931-4_20
13. Young S., Kershaw D., Odel J., Ollason D., Valtchev V., Woodland P. *The HTK Book*. Cambridge University Engineering Department, 2002, 271 p.
14. Schwarz P. *Phoneme Recognition Based on Long Temporal Context*. Ph.D. thesis. Brno University of Technology, 2008, 75 p.
15. Chernykh G., Korenevsky M., Levin K., Ponomareva I., Tomashenko N. State level control for acoustic model training. *Lecture Notes in Computer Science*, 2014, vol. 8773, pp. 435–442.
16. Chernykh G.A., Korenevsky M.L., Levin K.E., Ponomareva I.A., Tomashenko N.A. Krossvalidatsionnyi kontrol' sostoyanii pri obuchenii akusticheskikh modelei sistem avtomaticheskogo raspoznavaniya rechi [Cross-Validation State Control in Acoustic Model Training of Automatic Speech Recognition System]. *Izv. vuzov. Priborostroenie*, 2014, vol. 57, no. 2, pp. 23–28.

17. Moreno P., Joerg C., Van Thong J.-M, Glickman O. A recursive algorithm for the forced alignment of very long audio segments. *Proc. Int. Conf. on Spoken Language Processing, ICSLP 1998*. Sydney, Australia, 1998, pp. 2711–2714.
18. Khokhlov Y., Tomashenko N. Speech recognition performance evaluation for LVCSR system. *Proc. 14th Int. Conf. on Speech and Computer, SPECOM 2011*. Kazan', Russia, 2011, pp. 129–135.
19. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, vol. 27, no. 8, pp. 861–874. doi: 10.1016/j.patrec.2005.10.010

Булгакова Елена Владимировна – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, bulgakova@speechpro.com

Шолохов Алексей Владимирович – аспирант, Университет Восточной Финляндии, Йоэнсуу, FI-80101, Финляндия, sholohov@speechpro.com

Томашенко Наталья Александровна – младший научный сотрудник, ООО «Центр речевых технологий», Санкт-Петербург, 196084, Российская Федерация; инженер, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, tomashenko-n@speechpro.com

Elena V. Bulgakova – postgraguate, ITMO University, Saint Petersburg, 197101, Russian Federation, bulgakova@speechpro.com

Alexei V. Sholokhov – postgraduate, University of Eastern Finland, Joensuu, FI-80101, Finland, sholohov@speechpro.com

Natalia A. Tomashenko – junior scientific researcher, "Speech Technology Center", LLC, Saint Petersburg, 196084, Russian Federation; engineer, ITMO University, Saint Petersburg, 197101, Russian Federation, tomashenko-n@speechpro.com