НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2016 Том 16 № 2 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2016 Vol. 16 No 2 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 621.391.037.372

ПОЛУАВТОМАТИЧЕСКАЯ СИСТЕМА ВЕРИФИКАЦИИ ДИКТОРОВ

Е.В. Булгакова[a], А.В. Шолохов[b]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] Университет Восточной Финляндии, Йоэнсуу, FI-80101, Финляндия
Адрес для переписки: bulgakova@speechpro.com

Информация о статье
Поступила в редакцию 16.12.15, принята к печати 26.01.16
doi:10.17586/2226-1494-2016-16-2-284-289
Язык статьи – русский

Ссылка для цитирования: Булгакова Е.В., Шолохов А.В. Полуавтоматическая система верификации дикторов // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 2. С. 284–289. doi:10.17586/2226-1494-2016-16-2-284-289

Аннотация

Предмет исследования. Представлена полуавтоматическая система верификации диктора по речи на основе сравнения значений формант, статистик длительностей звуков, а также мелодических характеристик. В последнее время благодаря развитию речевых технологий наблюдается значительный интерес к поиску экспертных систем верификации дикторов по голосу, обладающих высокой надежностью, а также низкой трудоемкостью за счет автоматизации процессов обработки данных для экспертного анализа. **Описание системы.** Впервые представлено описание системы, позволяющей анализировать сходство либо различие голосов дикторов на основе сравнения статистик длительностей фонем, формантных признаков и мелодических характеристик. Характерной особенностью предложенной системы, в основе которой лежит принцип фузирования (объединения) методов, является слабая корреляция между анализируемыми признаками, что приводит к общему снижению ошибки распознавания диктора. Преимуществом системы является возможность проведения экспресс-исследования фонограмм благодаря автоматизации процессов подготовки данных и принятия решения. Описываются принципы работы методов и способ их фузирования. **Основные результаты.** Проведена апробация системы на базе 1190 пар записей «свой–свой» и 10450 пар записей вида «свой–чужой». Записи включают русскую речь дикторов-мужчин и дикторов-женщин. Точность распознавания составила 98,59% для записей мужской речи и 96,17% для записей женской речи. Также было экспериментально установлено, что из всех используемых методов наиболее надежным является формантный метод. **Практическая значимость.** Результаты эксперимента показали применимость предложенной системы для решения задачи распознавания диктора по голосу и речи в рамках проведения фоноскопической экспертизы.

Ключевые слова
фоноскопическая экспертиза, распознавание диктора, полуавтоматические методы верификации, статистика длительностей фонем, формантные признаки, мелодические характеристики

SEMI-AUTOMATIC SPEAKER VERIFICATION SYSTEM

E.V. Bulgakova[a], A.V. Sholokhov[b]

[a] ITMO University, 197101, Saint Petersburg, Russian Federation
[b] University of Eastern Finland, Joensuu, FI-80101, Finland
Corresponding author: bulgakova@speechpro.com

Article info
Received 16.12.15, accepted 26.01.16
doi:10.17586/2226-1494-2016-16-2-284-289
Article in Russian

For citation: Bulgakova E.V., Sholokhov A.V. Semi-automatic speaker verification system. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 2, pp. 284–289. doi:10.17586/2226-1494-2016-16-2-284-289

Abstract

Subject of Research. The paper presents a semi-automatic speaker verification system based on comparing of formant values, statistics of phone lengths and melodic characteristics as well. Due to the development of speech technology, there is an increased interest now in searching for expert speaker verification systems, which have high reliability and low labour intensiveness because of the automation of data processing for the expert analysis. **System Description**. We present a description of a novel system analyzing similarity or distinction of speaker voices based on comparing statistics of phone lengths, formant features and melodic characteristics. The characteristic feature of the proposed system based on fusion of methods is a weak correlation between the analyzed features that leads to a decrease in the error rate of speaker recognition.

The system advantage is the possibility to carry out rapid analysis of recordings since the processes of data preprocessing and making decision are automated. We describe the functioning methods as well as fusion of methods to combine their decisions. **Main Results**. We have tested the system on the speech database of 1190 target trials and 10450 non-target trials, including the Russian speech of the male and female speakers. The recognition accuracy of the system is 98.59% on the database containing records of the male speech, and 96.17% on the database containing records of the female speech. It was also experimentally established that the formant method is the most reliable of all used methods. **Practical Significance**. Experimental results have shown that proposed system is applicable for the speaker recognition task in the course of phonoscopic examination.

Keywords
phonoscopic examination, speaker recognition, semi-automatic speaker verification methods, statistics of phone lengths, formant features, melodic characteristics

Введение

Речевой сигнал содержит различную информацию, в том числе индивидуальные голосовые характеристики, позволяющие узнать человека по голосу и, как следствие, решить задачу распознавания диктора. Данная задача условно делится на две подзадачи: верификацию диктора, в случае если необходимо дать бинарный ответ о тождестве либо различии голосов дикторов на эталонной и тестовой фонограммах, и идентификацию дикторов, в случае если необходимо из множества эталонных записей определить голос, тождественный голосу диктора на тестовой фонограмме. Разработанная система может быть использована для обеих подзадач, однако в рамках настоящей работы рассматривается задача верификации дикторов.

В настоящее время для решения проблемы распознавания диктора широко применяются как автоматические, так и экспертные методы. Использование экспертных методов в процессе проведения фоноскопических исследований с целью идентификации либо верификации говорящего дает возможность уточнить, скорректировать работу автоматических средств анализа и сравнения речевых сигналов. Однако применение данных методов ограничено необходимостью привлечения высококвалифицированных экспертов. Кроме того, экспертные методы обладают значительной трудоемкостью, что затрудняет их использование в условиях временных ограничений. Общее решение в результате применения экспертных методов во многом субъективно, поскольку зависит от личного опыта эксперта [1]. К числу недостатков экспертных систем анализа речевого сигнала следует также отнести наличие «адаптивных процедур», позволяющих вмешиваться в процедуру принятия решения, что приводит к увеличению влияния человеческого фактора в рамках проведения фоноскопического исследования [1]. Предложенная полуавтоматическая система верификации дикторов по голосу не обладает перечисленными выше недостатками. Данная система включает методы сравнения разных голосовых характеристик, которые были разработаны либо автоматизированы на основе экспертных методов. Так, метод распознавания диктора на основе сравнения статистик длительностей фонем был разработан [2] и в рамках данного исследования апробирован на большой базе с целью получения статистически достоверных результатов; метод сравнения мелодического контура [3] был автоматизирован на этапе подготовки данных для исследования, что также позволило оценить производительность данного метода на большой речевой базе; метод на основе сравнения формантных признаков был реализован на основе известного подхода [4], обладающего высокой точностью верификации, а также возможностью полной автоматизации процедуры сравнения. Обобщенное решение о сходстве либо различии голосов дикторов принимается автоматически, в результате фузирования используемых методов. Более ранние исследования в области криминалистического распознавания дикторов по голосу были сосредоточены на статистическом анализе распределения таких акустических и просодических признаков, как частота основного тона [3, 5, 6], частоты формант [7–9] и темпоральные супрасегментные характеристики [10, 11]. До недавнего времени относительно мало внимания уделялось исследованию специфичных для речи диктора статистик длительностей фонем. Однако подобная информация является значимой для решения задачи распознавания диктора, на что обращали внимание некоторые исследователи [2, 12]. Данные признаки особенно полезны для верификации дикторов с похожим строением речевого аппарата, когда некоторые другие признаки (например, спектральные характеристики) недостаточно надежны. В настоящей работе мы исследуем возможность использования длительностных характеристик совместно с другими признаками, не имеющими сильной корреляции друг с другом, для решения задачи верификации диктора [13]. Для этого мы проводим фузирование методов на основе длительностей фонем, сравнения формантных признаков, а также анализа мелодических характеристик. Таким образом, целью настоящего исследования является разработка полуавтоматической системы верификации дикторов по голосу на основе фузирования данных методов.

Методы верификации дикторов по голосу

Метод на основе сравнения формантных признаков. Положение областей усиленных частот (формант) в спектре речевого сигнала зависит от анатомической структуры голосового тракта и размеров резонансных полостей. По этой причине указанные спектральные характеристики могут быть

использованы для решения задачи распознавания дикторов. Из всех полуавтоматических методов, основанных на сравнении значений формант, был выбран и реализован подход, предложенный в [4], в связи с тем, что данный метод обладает высокой точностью верификации и низкой трудоемкостью, т.е. из всех «ручных» процедур подразумевает только полуавтоматическое выделение формантных треков. В наших экспериментах выделялись треки и определялись значения первых четырех формант для шести русских гласных. Далее к полученным значениям формантных треков был применен один из наиболее распространенных подходов к моделированию (аппроксимации) сложных многомерных распределений в задаче распознавания диктора – метод на основе СГР-УФМ [14]. Основная идея этого подхода заключается в построении так называемой универсальной фоновой модели (УФМ), которая аппроксимирует распределение признаков большого количества дикторов с целью полного представления всей популяции. Для создания универсальной фоновой модели обычно используются смеси гауссовых распределений (СГР).

Метод мелодического контура. Данный метод позволяет эксперту анализировать и сравнивать основные характеристики мелодических структур, представленные в виде наборов значений параметров основного тона (ОТ) для сопоставимых участков мелодического контура. 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. Анализ ОТ состоит в поиске одинаковых интонационных структур на рассматриваемых фонограммах и сравнении их характеристик. В качестве параметров ОТ были использованы следующие характеристики: минимум, максимум, средняя частота, интервал основного тона, измеренный в полутонах и герцах, скорость изменения тона, коэффициент изрезанности [3]. Обработка речевой базы данных включает полуавтоматическое получение графиков ОТ и их экспертную корректировку для утвердительных эмоционально нейтральных высказываний, а также подготовку таблиц, содержащих значения параметров ОТ для выделенных экспертом сопоставимых интонационных структур (длительных речевых фрагментов, синтагм, ядерных слогов и т.д.). В связи с высокой трудоемкостью данного метода с помощью специальных утилит была проведена сегментация речевого материала и подготовлены таблицы значений параметров ОТ только на основе длительных речевых фрагментов. Подобный подход возможен, так как данный интонационный фрагмент является наиболее информативным. Проведенная автоматизация подготовки данных для последующего анализа позволила в настоящей работе осуществить апробацию метода сравнения мелодического контура на большой речевой базе с целью получения статистически значимых результатов.

Метод на основе сравнения длительностей фонем. Основные этапы работы алгоритма, основанного на анализе статистик длительностей фонем, включают:

1. автоматическую фонемную сегментацию на основе фонограмм и их текстового содержания. Сегментация проводится в результате выравнивания (force alignment) транскрипции и звукового сигнала. Количество фонемных классов равно 53, что соответствует 52 фонам русского языка (17-ти гласным и 35-ти согласным) и модели паузы. Шесть символов гласных звуков (/i/, /e/, /a/, /u/, /o/, /y/) имеют числовой индекс, определяющий положение гласного по отношению к ударному слогу: «0» обозначает гласный в ударной позиции; «1» – гласный предударного слога; «2» определяет гласный /a/ второго предударного слога; «4» – любой заударный гласный. В процессе сегментации определяются временные границы каждого фона. После проведения фонемной сегментации эксперт может скорректировать границы выделенных фонов в случае необходимости. На рис. 1 представлен пример фонемной сегментации;
2. расчет средней длительности каждого фона, выделенного по фонемной разметке;
3. вычисление параметрической оценки сходства голосов дикторов и принятие решения.

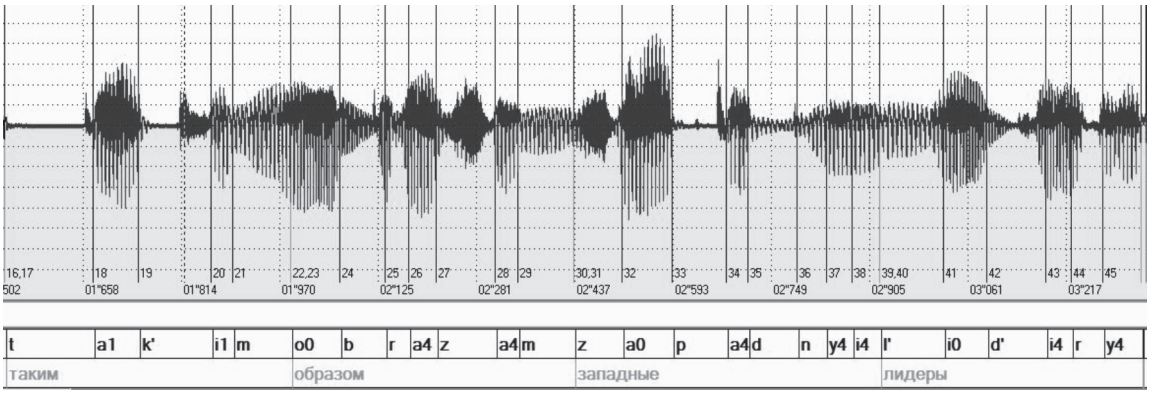

Рис. 1. Пример фонетической сегментации высказывания «takim obrazom zapadnye lidery»

Более подробное описание алгоритма представлено в [2], где он был протестирован на речевой базе, содержащей лишь незначительное количество записей. В данной работе проведена апробация метода на большой речевой базе, описанной в разделе «Экспериментальные результаты».

Фузирование методов верификации дикторов

Для повышения надежности систем верификации дикторов может приниматься общее решение, полученное на основе решений отдельных методов [15]. Такая процедура называется объединением или фузированием методов на уровне оценок сходства. Существуют различные способы расчета общего решения, однако наиболее распространенным из них является линейная комбинация оценок сходства. Для оценок сходства s_i, полученных от нескольких методов, результирующая оценка рассчитывается по следующей формуле:

$$s = \sum_i a_i s_i \quad \text{при} \sum_i a_i = 1,$$

где суммирование ведется по всем фузируемым методам, а a_i – неотрицательные веса, которые можно интерпретировать как «вклад» или «важность» того или иного метода в финальную оценку сходства. Оптимальные значения весов при этом оцениваются на отдельной выборке.

Важным аспектом фузирования является требование статистической независимости оценок сходства между объединяемыми методами. В противном случае общее решение может не принести прироста в точности верификации дикторов.

В следующем разделе представлены эксперименты по сравнению точности работы систем верификации дикторов, основанных на методах, описанных выше, а также системы, построенной на их объединении. Приведены результаты экспериментов по сравнению точности работы систем верификации дикторов на основе фузирования этих методов.

Экспериментальные результаты

Для оценки параметров алгоритмов была использована речевая база данных 194 носителей русского языка. База содержит квазиспонтанную русскую речь 124 дикторов-мужчин и 70 дикторов-женщин, записанных через телефонный канал. Во время записи каждый информант отвечает на заранее подготовленные вопросы. Каждый диктор принимает участие в пяти сессиях записи длительностью от 3 до 5 минут с интервалом 1 неделя между сессиями. Для тестирования была записана база спонтанных телефонных диалогов между носителями русского языка длительностью 1–3 минуты. Данный тестовый материал состоит из 773 пар фонограмм вида «свой–свой» и 8394 пар фонограмм «свой–чужой» для мужчин, а также 417 пар записей «свой–свой» и 2056 пар записей «свой–чужой» для женщин.

Далее мы рассмотрим эксперименты по верификации дикторов с использованием базы, описанной выше. Для оценки эффективности результатов верификации дикторов использовалось значение равенства ошибок I и II рода – равновероятная ошибка (Equal Error Rate, EER, %) [16]. В рамках эксперимента мы сравнили точность верификации каждого из трех описанных методов. Следует отметить, что все три метода были протестированы в автоматическом режиме, т.е. без правки формантных треков, границ фонов и графиков основного тона.

В таблице представлены результаты сравнения. Как следует из данных, приведенных в таблице, формантный метод является наиболее надежным.

Метод	мужчины	женщины
Основной тон	23, 28	27,33
Длительность фонем	27,57	36,98
Форманты	2,93	4,63
Форманты + длительности фонем	2,02	4,49
Форманты + длительность фонем + основной тон	1,41	3,83

Таблица. Показатели EER-верификации по речи для двух полов дикторов, %

Чтобы изучить возможность совместного использования сравниваемых методов, мы провели их фузирование. В предыдущих исследованиях было выполнено фузирование методов на основе статистик длительностей фонем и формантных признаков [17]. В настоящей работе мы осуществили фузирование всех трех методов.

Результаты, приведенные в таблице, показывают, что фузирование методов на основе слабо коррелированных признаков (мелодических характеристик, формантных признаков, длительностей фонем) приводит к снижению EER и повышает точность работы системы распознавания дикторов.

Обсуждение результатов

Как показывают результаты проведенного эксперимента, разработанная система может быть применима для решения задачи распознавания диктора по речи в условиях временных ограничений в рамках проведения фоноскопической экспертизы в связи с низкой трудоемкостью проведения исследования, а также по той причине, что ошибка верификации данной системы сопоставима с уровнем ошибки автоматических методов [18]. Система была апробирована на записях, содержащих русскоязычный материал, и может быть рекомендована экспертам-фоноскопистам.

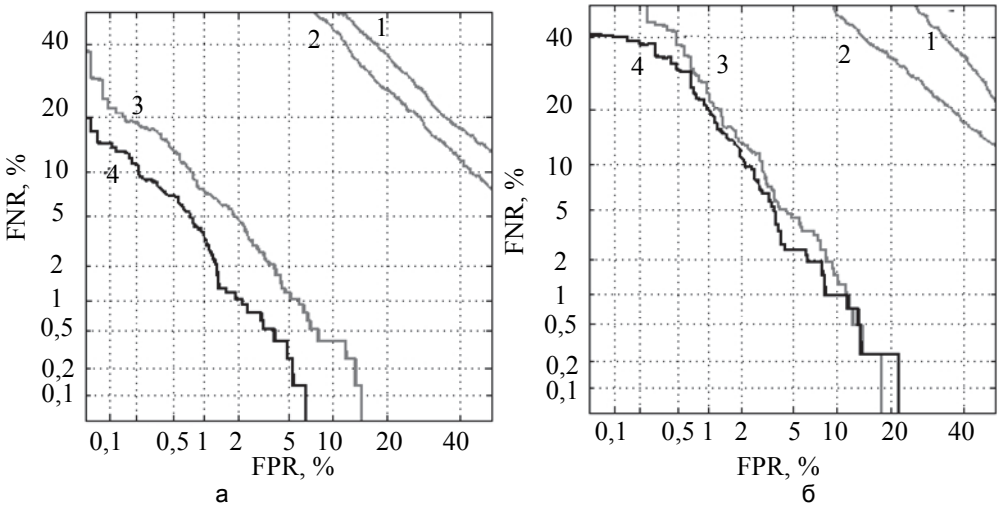

Рис. 2. DET-кривые для мужчин (а) и женщин (б). Кривая 1 демонстрирует производительность формантного метода, кривая 2 – метода основного тона, кривая 3 – метода длительностей фонем, кривая 4 показывает производительность системы в целом. FNR (False Negative Rate), FPR (False Positive Rate)

Полученные данные (рис. 2) подтверждают экспериментально установленный вывод [18] о том, что при верификации женских голосов алгоритмы, как правило, показывают худшую эффективность, чем в случае с мужскими голосами.

Заключение

В работе впервые предложена полуавтоматическая система верификации дикторов на основе сравнения признаков, не имеющих ярко выраженной корреляции друг с другом: статистик длительностей фонем, характерных для речи конкретного диктора, формантных признаков и мелодических характеристик. В рамках данного исследования проведены эксперименты, сделаны выводы и обозначены перспективы применения системы.

Преимуществом системы является возможность ее использования для проведения экспресс-анализа фонограмм. Точность работы системы составляет 98,59% на базе, содержащей записи мужской речи, и 96,17% на базе, содержащей записи женской речи.

References

1. Galyashina E.I. Linguistic analysis in the speaker identification systems: integrated complex examination approach based on forensic science technology. *Computational Linguistics and Intellectual Technologies*, 2015, vol. 1, pp.156–159.
2. 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). doi: 10.17586/2226-1494-2015-15-1-70-77
3. Smirnova N., Starshinov A., Goloshchapova T., Oparin I. 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.
4. Becker T., Jessen M., Grigoras C. Forensic speaker verification using formant features and Gaussian mixture models. *Proc. 9th Annual Conference of the International Speech Communication, INTERSPEECH 2008*. Brisbane, Australia, 2008, pp. 1505–1508.
5. Kunzel H.J., Masthoff H.R., Koster J.P. The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition. *Science and Justice*, 1995, vol. 35, no. 4, pp. 291–295. doi: 10.1016/S1355-0306(95)72683-1

6. Nolan F. Intonation in speaker identification: an experiment on pitch alignment features. *Speech, Language and the Law*, 2002, vol. 9, no. 1, pp. 1–21.
7. Morrison G.S. Likelihood-ratio-based forensic speaker comparison using representations of vowel formant trajectories. *Journal of the Acoustical Society of America*, 2009, vol. 125, pp. 2387–2397. doi: 10.1121/1.3081384
8. Nolan F., Grigoras C. A case for formant analysis in forensic speaker identification. *International Journal of Speech Language and the Law*, 2005, vol. 12, no. 2, pp. 143–173. doi: 10.1558/sll.2005.12.2.143
9. Rose P., Osanai T., Kinoshita Y. Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Speech Language and the Law*, 2003, vol. 10, no. 2, pp. 179–202.
10. Dellwo V., Leemann A., Kolly M.-J. Speaker idiosyncratic rhythmic features in the speech signal. *Proc. 13th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2012. Portland, USA, 2012, pp. 1582–1585.
11. Leemann A., Kolly M.-J., Dellwo V. Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Science International*, 2014, vol. 238, pp. 59–67. doi: 10.1016/j.forsciint.2014.02.019
12. Van Heerden C., Barnard E. Speaker-specific variability of phoneme durations. *South African Computer Journal*, 2008, vol. 40, pp. 44–50.
13. Matveev Y.N. Study of informative speech features for automatic speaker identification. *Journal of Instrument Engineering*, 2013, vol. 56, no. 2, pp. 47–51.
14. Reynolds D.A., Quatieri T.E., Dunn R.B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, vol. 10, no. 1, pp. 19–41. doi: 10.1006/dspr.1999.0361
15. Matveev Y.N. Evaluation of the confidence interval for decision prediction of an ensemble of classifiers. *Journal of Instrument Engineering*, 2013, vol. 56, no. 2, pp. 74–79.
16. The NIST year 2010 Speaker Recognition Evaluation plan. Available at: http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf (accessed 02.02.2016).
17. Bulgakova E., Sholohov A., Tomashenko N., Matveev Y. Speaker verification using spectral and durational segmental characteristics. *Lecture Notes in Computer Science*, 2015, vol. 9319, pp. 397–404. doi: 10.1007/978-3-319-23132-7_49
18. 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

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

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

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