АКУСТИЧЕСКИЙ ЖУРНАЛ, 2024, том 70, № 1, с. 113–119

АКУСТИКА ЖИВЫХ СИСТЕМ. БИОМЕДИЦИНСКАЯ АКУСТИКА

УДК 534.4,534.7,534.8

РАСПОЗНАВАНИЕ ЛИЧНОСТИ ПО ГОЛОСУ НА БАЗЕ ПАРАМЕТРОВ СПЕКТРАЛЬНОЙ МОДЕЛИ ГОЛОСОВОГО ИСТОЧНИКА

© 2024 г. И. С. Макаров[a], *, Д. С. Осипов[b, c], **

[a]ООО "БиометрикЛабс", 4-я ул. 8-го Марта 3, стр. 3, Москва, 125319 Россия

[b]Институт проблем передачи информации им. А.А. Харкевича РАН, Большой Каретный переулок 19, стр. 1, Москва, 127051 Россия

[c]Национальный исследовательский университет "Высшая школа экономики", ул. Мясницкая 20, Москва, 101000 Россия

**e-mail: im@biometriclabs.ru*

***e-mail: d_osipov@iitp.ru, dosipov@hse.ru*

Поступила в редакцию 15.02.2023 г.
После доработки 07.07.2023 г.
Принята к публикации 19.09.2023 г.

Исследована информативность параметров спектральной модели голосового источника в задаче автоматического распознавания личности по голосу. Для голосовых параметров ошибка распознавания личности составила 20.8%; совместное использование этих параметров с периодом основного тона понизило ошибку до 13.8%. Наконец, совместное использование параметров спектральной модели с периодом основного тона и мел-частотными кепстральными коэффициентами обеспечило наивысшую точность (ошибка распознавания составила 1.2%).

Ключевые слова: обратная фильтрация, голосовой источник, математические модели голосообразования, распознавание личности по голосу

DOI: 10.31857/S0320791924010132 **EDN**: ZNOOAD

ВВЕДЕНИЕ

Вопрос об информативности параметров голосового источника в задаче автоматического распознавания личности по голосу является дискуссионным. В то время как высокая информативность характеристик передаточной функции речевого тракта в системах голосовой идентификации/верификации личности несомненна [1], полезность голосового источника для биометрических приложений неочевидна, а результаты разных публикаций зачастую противоречивы. Так, в [2] сделан вывод о достаточно высокой информативности характеристик голосового возбуждения в задаче распознавания личности по голосу (процент равновероятных ошибок 1-го и 2-го рода Equal Error Rate (EER) составил порядка 14%). Напротив, по мнению [3], голосовой источник практически не содержит индивидуальной информации о голосе человека (EER около 40%).

Отмеченное несовпадение результатов, полученных авторами разных работ, объясняется использованием разных алгоритмов определения и анализа голосового источника, а также разных речевых баз для обучения и тестирования систем распознавания личности. Так, в [4] источник аппроксимируется некоторой математической моделью первой производной голосового возбуждения, а в качестве речевой базы используется акустическая база TIMIT (600+ дикторов, записанных в студийных условиях с отношением сигнал/шум >25 дБ и частотой дискретизации 16 кГц). В [5] голосовое возбуждение параметризуется набором

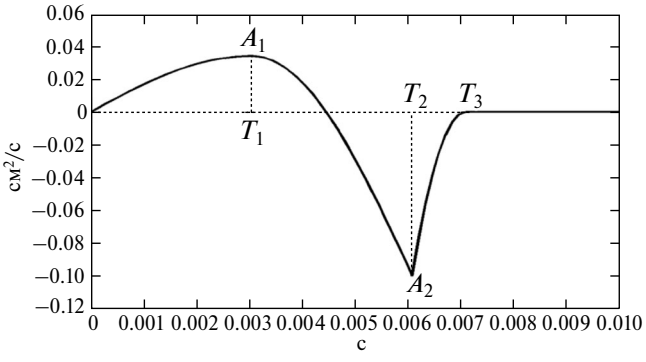

Рис. 1. Один период голосового источника, порождаемый моделью из [9]. T_0 = 10 мс (частота основного тона 100 Гц).

мел-частотных кепстральных коэффициентов, а в качестве речевой базы используется TIMIT и YOHO (100+ дикторов, записанных через телефонный канал с частотой дискретизации 8 кГц). В [2] в качестве параметров голосового источника выступают отклики одного из внутренних слоев авто-ассоциативной нейронной сети (автоэнкодера), а обучение и тестирование системы распознавания личности по голосу осуществляется с помощью базы из 30-ти дикторов, записанных через микрофон (16 кГц) и телефон (8 кГц) при отношении сигнал/шум <20 дБ. Авторы [3, 6, 7] в качестве речевых баз используют базы NIST SRE 2006 и 2010 (телефонный канал) и параметризуют голосовой источник набором мел-частотных кепстральных коэффициентов. В [8] по речевому сигналу вычисляется спектр Фурье голосового возбуждения путем нахождения отношения спектров на открытой и закрытой фазах голосовой щели; несколько коэффициентов при главных компонентах, аппроксимирующих спектр голосового источника, применяются для автоматического распознавания личности по голосу. В качестве речевых баз используется а) база записей 200+ дикторов через четыре типа микрофонов (16 кГц), б) та же база, но пропущенная через низкоскоростной кодек, в) база TIMIT.

Задача настоящей статьи – исследовать вопрос об информативности параметров некоторой модели голосового источника в задаче автоматического распознавания личности по голосу. В качестве модели источника выбрана модель, построенная в [9]. В качестве метода распознавания личности по голосу использован подход на базе *i*-векторов, который долгое время (примерно до 2019 г.) являлся *state of the art* в голосовой биометрии. Хотя в настоящее время подход существенно потеснен системами, использующими так называемые глубокие признаки (т.е. отклики слоев глубокой нейронной сети определенной топологии, например, *x*-векторы [10]), точность систем на базе *i*-векторов до сих пор используется в качестве референтной при исследовании и сравнении алгоритмов на базе глубоких нейронных сетей [11]. Кроме того, комбинация *i*-векторов с глубокими признаками, как правило, повышает точность общей системы [12, 13].

Структура статьи следующая. Раздел 1 посвящен построению спектральной модели голосового источника. В Разделе 2 описываются эксперименты по автоматическому распознаванию личности на базе параметров построенной модели и обсуждаются полученные результаты.

СПЕКТРАЛЬНАЯ МОДЕЛЬ ГОЛОСОВОГО ИСТОЧНИКА

В основе дальнейшего изложения – модель голосового источника, построенная в [9] и исследованная в [14, 16]. Выбор модели продиктован, с одной стороны, ее вычислительной простотой, а с другой стороны, ее эффективностью при решении ряда практических задач речевых технологий: задачи определения пола человека по голосу [14], голосовой идентификации личности [15] и обратной фильтрации речевого сигнала [16].

Исходная модель сформулирована во временной области и описывает форму голосового источника как функцию времени t на одном периоде основного тона T_0. Модель определяется следующими соотношениями:

$$W = \begin{cases} W_1 = A_1 \sin\left(\dfrac{\pi t}{2T_1}\right), 0 \le t \le T_1, \\ W_2 = \left(A_1 + A_2\right)\cos\left(\dfrac{\pi\left(t - T_1\right)}{2\left(T_2 - T_1\right)}\right) - \\ -A_2, T_1 < t \le T_2, \\ W_3 = -A_2\left(\dfrac{T_3 - t}{T_3 - T_2}\right)^2, T_2 < t \le T_3, \\ 0, T_3 < t \le T_0. \end{cases} \quad (1)$$

Форма голосового источника, порождаемая данной моделью, для $T_0 = 10$ мс (т.е. для частоты основного тона 100 Гц) показана на рис. 1. На том же рисунке показаны параметры, использованные в соотношении (1): A_1 – максимальная амплитуда голосового импульса, A_2 – минимальная амплитуда голосового импульса (= амплитуда голосового возбуждения), T_1 – момент времени, соответствующий максимуму объемной скорости воздушного потока, протекающего через голосовую щель, T_2 – момент времени, соответствующий моменту схлопывания голосовых складок, T_3 – момент времени, соответствующий началу фазы сомкнутых голосовых складок.

Соотношения (1) зависят от пяти независимых параметров A_2, T_0, T_1, T_2, T_3, а шестой параметр A_1 является функцией остальных:

$$A_1 = \frac{A_2}{6T_2}\left[(6 - 3\pi)T_1 + (2\pi - 6)T_2 + \pi T_3\right]. \quad (2)$$

Основной недостаток исходной модели заключается в необходимости автоматически оценивать по речевому сигналу моменты времени T_1, T_2, T_3 для аппроксимации голосового возбуждения. Такая оценка нередко затруднительна из-за искажений речевого сигнала различными помехами и индивидуальных особенностей голоса человека [17]. Чтобы уйти от необходимости оценки этих параметров, переведем модель (1)–(2) в частотную область. Предполагая функцию W квази-периодической и разлагая ее в ряд Фурье, получим значение

n-ной комплексной амплитуды спектра модели: $c_n = S_1 + S_2 + S_3$, где:

$$S_1 = \frac{A_1}{\left(\frac{\pi^2}{4T_1^2} - n^2\omega_0^2\right)T_0} \times \left(\frac{\pi}{2T_1} - jn\omega_0 \exp\left(-jn\omega_0 T_1\right)\right), \quad (3а)$$

$$S_2 = \frac{A_1 + A_2}{\left(\frac{\pi^2}{4(T_2 - T_1)^2} - n^2\omega_0^2\right)T_0} \times \left(\frac{\pi \exp\left(-jn\omega_0 T_2\right)}{2(T_2 - T_1)} + jn\omega_0 \exp\left(-jn\omega_0 T_1\right)\right) + \frac{A_2}{jn\omega_0 T_0}\left(\exp\left(-jn\omega_0 T_2\right) - \exp\left(-jn\omega_0 T_1\right)\right), \quad (3б)$$

$$S_3 = -\frac{A_2}{T_0(T_3 - T_2)^2} \times \left[\begin{array}{c} \frac{2}{jn^3\omega_0^3}\exp\left(-jn\omega_0 T_3\right) + \frac{(T_3 - T_2)^2}{jn\omega_0} \times \\ \times \exp\left(-jn\omega_0 T_2\right) + \frac{2(T_3 - T_2)}{n^2\omega_0^2} \times \\ \times \exp\left(-jn\omega_0 T_2\right) - \frac{2}{jn^3\omega_0^3}\exp\left(-jn\omega_0 T_2\right) \end{array}\right]. \quad (3в)$$

Здесь $j = \sqrt{-1}$, $\omega_0 = 2\pi / T_0$.

На рис. 2 показано абсолютное значение спектра модели (3) для функции на рис. 1, вычисленного на сетке из 9-ти гармоник на частотах 200, 300, ..., 1000 Гц с основной гармоникой на частоте 100 Гц.

Комплексная амплитуда c_n n-ной гармоники спектральной модели (3) зависит от пяти параметров: T_0, A_2, T_1, T_2 и T_3. Параметр T_0, соответствующий периоду основного тона, определялся непосредственно по речевому сигналу с помощью алгоритма из [18]. Остальные параметры вычислялись следующим образом:

1. Для каждого вокализованного участка входного речевого сигнала определялась функция сигнала-остатка с помощью процедуры линейного предсказания [14].

2. Сигнал-остаток разбивался на последовательность перекрывающихся временных фреймов (мы использовали окно Хэмминга длительностью 20 мс с перекрытием 10 мс).

3. Для каждого временного фрейма оценивался соответствующий ему период основного тона.

4. Для каждого фрейма сигнала-остатка вычислялся соответствующий ему спектр Фурье (мы использовали 1024-точечное Быстрое Преобразование Фурье).

5. По известной частоте основного тона определялся модуль M_0 амплитуды основной гармоники, а также модули M_1, ..., M_N амплитуд N высших гармоник (во всех наших экспериментах $N = 9$) для каждого фрейма сигнала-остатка.

6. Для каждого фрейма параметры A_2, T_1, T_2 и T_3 подбирались путем минимизации невязки в евклидовой метрике $\|\bullet\|_2^2$ между амплитудами M_0, M_1, ..., M_N и амплитудами $\tilde{M}_0$, $\tilde{M}_1$, ..., $\tilde{M}_N$, вычисленными по модели (3). Минимизация осуществлялась методом градиентного спуска с ограничениями на искомые параметры: $0.05 \leqslant A_2 \leqslant 0.3$; $0.1 \leqslant T_1 \leqslant T_2 - 0.1$; $T_2 \leqslant T_3 - 0.1$; $T_3 \leqslant T_0$. Поскольку описанная задача минимизации является многоэкстремальной [17], для получения удовлетворительной точности аппроксимации спектра голосового источника моделью (3) необходим удачный выбор начальных приближений. Для решения этой задачи нами была составлена база данных, содержащая различные значения параметров A_2, T_1, T_2 и T_3 (всего около 10000 векторов). Эта база была подвергнута процедуре кластеризации методом K-средних и разбита на 16 кластеров. Процедура минимизации последовательно запускалась для каждого из 16-ти центроидов, после чего в качестве окончательного решения брались такие значения параметров A_2, T_1, T_2 и T_3, которые обеспечивали глобальный минимум среди всех итераций.

2. РАСПОЗНАВАНИЕ ЛИЧНОСТИ ПО ПАРАМЕТРАМ СПЕКТРАЛЬНОЙ МОДЕЛИ

Для исследования информативности параметров спектральной модели голосового источника в задаче распознавания личности по голосу была использована база записей русских цифр (от нуля до девяти) в произнесении 200 носителей русского языка с частотой дискретизации 16 кГц. При сборе базы каждая цифра произносилась каждым человеком 20 раз через пять параллельно расположенных микрофонов, имеющих существенно различные диаграммы направленности (направленный микрофон, кардиоидный микрофон, головная гарнитура с шумоподавлением, репортерский всенаправленный микрофон и микрофон сотового телефона NOKIA). Запись осуществлялась в две сессии – по 10 произнесений каждой цифры во все микрофоны в первую сессию и по 10 произнесений – во вторую. Перерыв между сессиями в среднем составлял 7 дней. Общее количество собранных файлов составило 200000 (200 человек ×

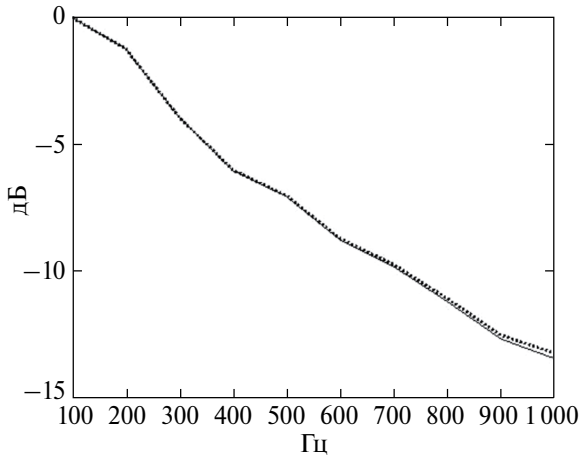

Рис. 2. Модуль спектра источника (рис. 1), вычисленного по модели (3) (сплошная линия) и с помощью 1024-точечного Быстрого Преобразования Фурье (пунктир).

× 10 цифр × 20 произнесений × 5 микрофонов). Отметим, что объем собранной базы выше, чем объем базы VoxCeleb 1 (= 153516 файлов), широко используемой для обучения и тестирования алгоритмов распознавания личности по голосу [19].

100 человек были использованы для обучения классификатора, 20 человек – для валидации параметров классификатора и 80 человек – для тестирования. Общий объем обучающей базы составил 100000 файлов (100 человек × 10 цифр × 20 произнесений × 5 микрофонов). Общее количество акустических фреймов обучающей выборки, по которым определялись необходимые акустические векторы, составил около 30 млн сегментов. В качестве акустических векторов использовались следующие наборы: а) параметры A_2, T_1, T_2 и T_3, б) те же параметры + T_0, в) 25 мел-частотных кепстральных коэффициентов вместе с их первыми производными и нулевым коэффициентом, г) параметры а) + в), д) параметры б) + в). Распознавание личности по голосу осуществлялось в рамках подхода на базе *i*-векторов [20], которые вычислялись по входным сигналам и затем сравнивались друг с другом в PLDA-метрике [21]. Универсальная модель (Universal Background Model, UBM), построенная на базе модели из 1024-х Гауссовых смесей с диагональными ковариационными матрицами, а также соответствующая матрица полной изменчивости обучались на данных всех ста дикторов из обучающей выборки. Предварительная кластеризация исходного акустического пространства осуществлялась с помощью алгоритма K-средних (10 итераций); дальнейшее уточнение параметров Гауссовых смесей (определение взвешивающих коэффициентов, средних векторов и векторов дисперсий) выполнялось с помощью EM-алгоритма (Expectation-Maximization) на 5-ти итерациях [1]. При вычислении матрицы полной изменчивости использовался алгоритм из [22], при этом число столбцов этой матрицы полагалось равным 800. Соответствующие 800-размерные *i*-векторы усекались до размерности 450 с помощью вероятностного линейного дискриминантного анализа (Probabilistic Linear Discriminant Analysis, PLDA). Отклики классификаторов по разным параметрам подвергались процедуре калибровки, после чего объединялись с помощью процедуры логистической регрессии [23].

Для каждого человека из тестовой выборки были сгенерированы 40 парольных фраз, где каждая парольная фраза представляла собой случайным образом склеенные цифры (после удаления неречевых сегментов) в произнесении этого человека. При склейке не учитывался тип микрофона, через который данная цифра была записана (иначе говоря, в одну и ту же парольную фразу могли попасть цифры, записанные с различных микрофонов). Средняя длительность парольных фраз по всей базе для всех дикторов составила примерно 4 с (от двух до 4-х произнесений цифр по каждой парольной фразе). В процессе тестирования на вход системы подавалась пара парольных фраз (эти фразы могли быть произнесены одним и тем же диктором либо двумя различными дикторами). Для каждой пары система возвращала либо индекс “0” (по мнению системы, обе фразы произнесены одним и тем же диктором), либо “1” (по мнению системы, фразы произнесены различными дикторами). Поскольку отклики классификаторов по всем параметрам были откалиброваны, во всех случаях порог принятия решений был равен нулю. По полученным оценкам вычислялись ошибки 1-го и 2-го рода, а по ним – EER (Equal Error Rate, процент равновероятных ошибок). Общее число пар для тестирования было равно $(80 \times 40)^2 = 10.24$ млн пар. Никакой предварительной настройки (enrollment) системы на голоса пользователей из тестовой базы не осуществлялось. Результаты тестов указаны в таблице.

Видно, что сами по себе параметры голосового источника ожидаемо не очень информативны (EER = 20.8%); совместное их использование с периодом основного тона T_0 понижает EER до 13.8%. Наконец, совместное использование этих параметров с мел-частотными кепстральными коэффициентами дает наивысшую точность (EER = 1.2%).

Представляет интерес сопоставление полученных результатов с результатами других исследователей, занимавшихся вопросом информативности параметров голосового источника в задаче голосовой биометрии.

В [4, 5] ошибка распознавания личности по параметрам голосового источника на базе TIMIT

составила 28.6 и 12.94% соответственно. Поскольку эта база имеет открытый доступ, мы провели на ней эксперименты по распознаванию личности с использованием двух наборов голосовых параметров: а) параметры A_2, T_1, T_2 и T_3, б) те же параметры + T_0 (при этом никакого дообучения наших моделей на базе TIMIT не осуществлялось – в экспериментах использованы только те модели, которые были обучены на 100 дикторах из нашей исходной акустической базы). Для набора а) ошибка распознавания составила 12.1%, а для б) 8.4%. Таким образом, на базе TIMIT наши модели голосового источника продемонстрировали значительно более высокую точность распознавания личности, чем модели, построенные в работах [4, 5]. Следует отметить, что база TIMIT записана в очень благоприятных с акустической точки зрения условиях (студийная запись через высококачественный микрофон с шумоподавлением, отношение сигнал/шум >25 дБ). Поэтому точность распознавания личности, полученная на этой базе, заведомо выше, чем точность на речевых базах, записанных в более реалистичных акустических условиях.

В [8] на речевой базе произнесений русских цифр, записанной через четыре типа микрофонов, ошибка распознавания мужчин по параметрам голосового источника составила 8%, а женщин – 15%. На первый взгляд, это очень хороший результат. Вместе с тем, в той же статье обученные на этой базе модели продемонстрировали на базе TIMIT крайне низкую точность распознавания (ошибка порядка 44%). Для TIMIT это худший результат среди всех известных работ, посвященных распознаванию по параметрам голосового источника. Это очень странно, если принять во внимание утверждение автора [8] о том, что исходная модель была обучена на акустически гораздо более разнообразной базе, чем TIMIT. На наш взгляд, это объясняется неустойчивостью алгоритмов из [8] относительно условий записи речевых сигналов; высокая точность распознавания личности, достигнутая на исходной базе, связана либо с переобучением моделей, либо с некорректно проведенными экспериментами. Попутно отметим, что в [8] делается ошибочное утверждение о том, что точность современных систем распознавания личности по голосу не превышает 85–90%. Это, разумеется, неверно: некоторые современные системы голосовой биометрии на базе VoxCeleb 1 демонстрируют точность ⩾99% [24]. Кроме того, в данной статье делается совершенно некорректное сопоставление точности тексто-зависимой системы голосовой биометрии (составляющей, по утверждению [8], 99.9%), требующей для тестирования обязательной предварительной настройки на диктора длительностью более 10 мин и парольных фраз длительностью десятки секунд, с точностью современных тексто-независимых систем, не предполагающих никакой настройки на диктора и работающих с фразами длительностью несколько секунд.

Таблица. Точность системы распознавания личности по голосу (EER, %)

Параметры	EER, %
A_2, T_1, T_2, T_3	20.8%
A_2, T_1, T_2, T_3 + T_0	13.8%
25 MFCC + 25 deltas + нулевой коэффициент	2.0%
A_2, T_1, T_2, T_3 + 25 MFCC + 25 deltas + нулевой коэффициент	1.6%
A_2, T_1, T_2, T_3 + T_0 + 25 MFCC + 25 deltas + нулевой коэффициент	1.2%

В работах [3, 6, 7] обучение и тестирование систем распознавания личности по параметрам голосового источника осуществлялось на базах NIST SRE 2006 и 2010. В [3] ошибка распознавания (измеренная как EER) составила 40%. Ошибка распознавания личности только по параметрам голосового источника в [6, 7] не приведена; вместе с тем, указано, что использование параметров источника наряду с мел-частотными кепстральными коэффициентами повысило точность распознавания на 3%. Прямое сопоставление этих результатов с нашими крайне затруднительно в связи с тем, что базы NIST SRE 2006 и 2010 – это телефонный канал с частотой дискретизации 8 кГц, а наша база – микрофонный канал с частотой 16 кГц. К сожалению, эти базы не имеют открытого доступа, поэтому мы не могли проверить на них наши модели. Отметим только, что и в наших экспериментах, и в экспериментах [3, 6, 7] обнаруживается одна и та же тенденция: добавление параметров источника к мел-частотным кепстральным коэффициентам приводит к повышению точности распознавания личности.

В [2] ошибка распознавания личности по параметрам голосового источника на микрофонном канале с частотой дискретизации 16 кГц составила 13.57%. Эта ошибка ниже, чем ошибка, полученная нами для параметров A_2, T_1, T_2 и T_3, (20.8%), и сопоставима с ошибкой распознавания для параметров T_0, A_2, T_1, T_2 и T_3 (13.8%). Отметим, что речевая база, использованная в [2], записана с одного микрофона, при этом в записи участвовало только 30 человек. Поскольку наша база существенно более разнообразна (как с точки зрения числа дикторов, так и с точки зрения акустических условий записи), мы полагаем, что

полученные нами оценки точности распознавания по параметрам голосового источника более достоверны.

Задача настоящей статьи – исследование вопроса информативности параметров голосового источника в распознавании личности по голосу. Авторы не ставили перед собой цели достичь более высокой точности распознавания, чем точность, демонстрируемая некоторыми современными нейросетевыми системами голосовой биометрии. Вместе с тем кратко остановимся на этом вопросе. Современные системы голосовой биометрии, построенные на базе глубоких нейронных сетей различных топологий, демонстрируют впечатляющую точность распознавания личности по голосу на акустически очень богатых выборках, содержащих голоса десятков и даже сотен тысяч человек, записанных в различных акустических условиях. Например, авторы [24] достигли точности распознавания 99% на базе VoxCeleb 1. Следует отметить, что существенное снижение ошибки распознавания в таких системах достигнуто не за счет использования определенных наборов акустических признаков, а за счет оптимизации топологии нейронных сетей, выбора специальной метрики для сравнения откликов скрытых слоев или же применения особых скрытых слоев (например, слоя внимания). По нашим результатам, использование параметров голосового источника в дополнение к стандартным мел-частотным кепстральным коэффициентам приводит к увеличению точности общей системы. Можно предположить, что встраивание модели голосового источника в существующие системы распознавания на базе глубоких нейронных сетей также может привести к снижению ошибки распознавания. Кроме того, представляет интерес построение классификатора параметров голосового источника не на базе *i*-векторов (как это сделано в настоящем исследовании), а на базе глубокой нейронной сети той или иной топологии (например, на базе *x*-признаков). Мы планируем проведение подобного исследования с дальнейшим объединением откликов по обоим классификаторам. Кроме того, мы планируем проведение тестов по исследованию информативности параметров голосового источника на базах VoxCeleb 1 и 2.

ЗАКЛЮЧЕНИЕ

Статья посвящена анализу информативности параметров голосового источника в задаче автоматического распознавания личности по голосу. В качестве модели голосового источника использовался спектральный вариант модели, построенной в [9]. Параметрами модели служили: амплитуда голосового возбуждения речевого тракта, а также некоторые параметры, характеризующие моменты открытия и закрытия голосовой щели. Распознавание личности по голосу осуществлялось в рамках парадигмы *i*-векторов; сравнение *i*-векторов друг с другом выполнялось в PLDA-метрике.

Для параметров спектральной модели голосового источника EER составил 20.8%; совместное использование этих параметров с периодом основного тона T_0 понизило EER до 13.8%. Наконец, совместное использование параметров спектральной модели с T_0 и мел-частотными кепстральными коэффициентами обеспечило наивысшую точность (EER = 1.2%).

В дальнейшем планируется провести эксперименты с параметрами спектральной модели голосового источника на открытых базах Vox Celeb 1 и 2 [19], а также использовать совместно с *i*-векторами классификаторы на базе глубоких нейронных сетей (напр., сравнение *x*-векторов в косинус-метрике) [10].

СПИСОК ЛИТЕРАТУРЫ

1. *Kinnunen T., Li H.* An overview of text-independent speaker recognition: From features to supervectors // Speech Commun. 2010. V. 52. P. 12–40.
2. *Yegnanarayana B., Mahadeva Prasanna S., Zachariah J., and Gupta Ch.* Combining Evidence from Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System // IEEE Trans. on Speech and Audio Process. 2005. V. 13. № 4. P. 575–582.
3. *Kinnunen T., Alku P.* On separating glottal source and vocal tract information in telephony speaker verification // Proc. the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2009.
4. *Plumpe M., Quatieri T., Reynolds D.* Modelling of the Glottal Flow Derivative Waveform with Application to Speaker Identification // IEEE Trans. on Speech and Audio Process. 1999. V. 7. № 5. P. 569–586.
5. *Gudnason J., Brookes M.* Voice Source Cepstrum Coefficients for Speaker Identification // IEEE ICASSP. 2008. P. 4821–4824.
6. *Mazaira-Fernández L., Álvarez-Marquina A., Gómez-Vilda P., Martínez Olalla R., Muñoz C.* Glottal Source Cepstrum Coefficients Applied to NIST SRE 2010 // V Jornadas de Reconocimiento Biométrico de Personas JRBP10, Huesca, España. 2010.
7. *Mazaira-Fernandes L., Alvares-Marquina A., Gomez-Vilda P.* Improving speaker recognition by biometric voice deconstruction // Front. Bioeng. Biotechnol. 2015. V. 3. P. 126.
8. *Sorokin V.N.* Vocal Source Contribution to Speaker Recognition // Pattern Recognition and Image Analysis. 2018. V. 28. № 3. P. 546–556.
9. *Ananthapadmanabha T.* Acoustic Analysis of Voice Source Dynamics // STL-QPSR. 1984. V. 2–3. P. 1–24.

10. *Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S.* X-vectors: Robust DNN Embeddings for Speaker Recognition // 2018 IEEE Int. Conf. on Speech, Audio, and signal Process. (ICASSP).
11. *Jessen M., Bortlik J., Schwarz P., Solewisz Y.* Evaluation of Phonexia Automatic Speaker Recognition Software under Conditions Reflecting Those of a Real Forensic Voice Comparison Case // Speech Communication 111. 2019. P. 22–28.
12. *Guo J., Xu N., Qian K., Shi Y., Wu Y., Alwan A.* Deep Neural Network based i-vector Mapping for Speaker Verification Using Short Utterances // Speech Communication 105. 2018. P. 92–102.
13. *Bai Z., Zhang X., Chen J.* Cosine Metric Learning based Speaker Verification // Speech Communication 118. 2020. P. 10–20.
14. *Сорокин В.Н., Макаров И.С.* Определение пола диктора по голосу // Акуст. журн. 2008. Т. 54. № 4. С. 659–668.
15. *Sorokin V.N., Leonov A.S.* Multisource Speech Analysis for Speaker Recognition // Pattern Recognition and Image Analysis. 2019. V. 29. № 1. P. 181–193.
16. *Леонов А.С., Сорокин В.Н.* Верхняя граница ошибок решения обратной задачи определения голосового источника // Акуст. журн. 2017. Т. 63. С. 532–545.
17. *Сорокин В.Н., Макаров И.С.* Обратная задача для голосового источника // Информационные процессы. 2006. Т. 6. № 4. С. 375–395.
18. *Цыплихин А.И.* Анализ импульсов голосового источника // Акуст. журн. 2007. Т. 53. № 1. С. 119–133.
19. *Nagrani A., Chung J.S., Xie W., Zisserman A.* Voxceleb: Large-scale speaker verification in the wild // Computer Science and Language, 2019.
20. *Dehak N., Kenny P., Dehak R., Dumouchel P., and Ouellet P.* Front-end factor analysis for speaker verification // IEEE Trans. on Audio, Speech, and Lang. Process. 2011. V. 19. № 4. P. 788–798.
21. *Kenny P., Stafylakis T., Ouellet P., Alam M., Dumouchel P.* PLDA for speaker recognition with utterances of arbitrary duration // Proc. ICASSP. 2013. P. 76449–7653.
22. *Vestman V., Kinnunen T.* Supervector Compression Strategies to Speed up i-Vector System Development // Speaker Odyssey 2018: The Speaker and Language Recognition Workshop.
23. *Morrison G.* Tutorial on logistic regression calibration and fusion: converting a score to a likelihood ratio // Australian Journal of Forensic Sciences. 2013. V. 45. № 2. P. 173–197.
24. *Zhu W., Kong T., Lu S., Li J., Zhang D., Deng F., Wang X., Yang S., Liu J.* SpeechNAS: Towards Better Trade-off between Latency and Accuracy for Large-Scale Speaker Verification // arXiv – CS – Artificial Intelligence, 2021.
https://doi.org/arxiv-2109.08839