Министерство образования Республики Беларусь
Учреждение образования
«Полоцкий государственный университет»

ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ: ДОСТИЖЕНИЯ, ПРОБЛЕМЫ, ИННОВАЦИИ (ИКТ-2018)

Электронный сборник статей
I Международной научно-практической конференции,
посвященной 50-летию Полоцкого государственного университета

(Новополоцк, 14–15 июня 2018 г.)

Новополоцк
Полоцкий государственный университет
2018

Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2018) [Электронный ресурс] : электронный сборник статей I международной научно-практической конференции, посвященной 50-летию Полоцкого государственного университета, Новополоцк, 14–15 июня 2018 г. / Полоцкий государственный университет. – Новополоцк, 2018. – 1 электрон. опт. диск (CD-ROM).

Представлены результаты новейших научных исследований, в области информационно-коммуникационных и интернет-технологий, а именно: методы и технологии математического и имитационного моделирования систем; автоматизация и управление производственными процессами; программная инженерия; тестирование и верификация программ; обработка сигналов, изображений и видео; защита информации и технологии информационной безопасности; электронный маркетинг; проблемы и инновационные технологии подготовки специалистов в данной области.

Сборник включен в Государственный регистр информационного ресурса. Регистрационное свидетельство № 3201815009 от 28.03.2018.

Компьютерный дизайн М. Э. Дистанова.

Технические редакторы: Т. А. Дарьянова, О. П. Михайлова.

Компьютерная верстка Д. М. Севастьяновой.

211440, ул. Блохина, 29, г. Новополоцк, Беларусь
тел. 8 (0214) 53-21-23, e-mail: irina.psu@gmail.com

ISBN 978-985-531-602-3

УДК 004.934.8’1

ДВУХУРОВНЕВАЯ АУТЕНТИФИКАЦИЯ ПОЛЬЗОВАТЕЛЯ ПО ГОЛОСОВОМУ СООБЩЕНИЮ

студент 4 курса А.С. ЯСКОВЕЦ, канд. физ.-мат. наук Е.И. КОЗЛОВА
(Белорусский государственный университет, Минск)

Речевые технологии в последнем десятилетии набирают все большую популярность. Многие из ведущих IT-компаний разработали собственные приложения в области распознавания речи, а с увеличением мощности мобильных устройств появилась возможность внедрения требовательных с точки зрения производительности решений «в карман» практически каждому человеку [3].

Далее будет представлена реализация системы текстозависимого распознавания человека по голосу. Процесс аутентификации можно разбить на две составляющие: идентификацию – определение соответствия поступающего на вход системы голоса одному из имеющихся в базе данных образцов, – и верификацию – подтверждение того, что голос является верно идентифицированным. Для этого необходимо создать две модели. Первая осуществляет идентификацию и обучается сразу на всей базе известных голосов, вторая проводит верификацию и обучается на каждом отдельно взятом голосе для достижения максимальной зависимости от него. Соответственно при успешной идентификации голоса парольная фраза, представляющая собой набор случайных цифр определенной длины, будет распознана верно – и второй этап успешно определит говорящего. При ошибке же на первом этапе модель, осуществляющая верификацию, будет обучена на неверном голосе, и вероятность ошибки, состоящей в ложноположительном срабатывании системы, при достаточной длине парольной фразы будет сведена к минимуму.

На рис. 1 представлено схематичное изображение двух этапов аутентификации.

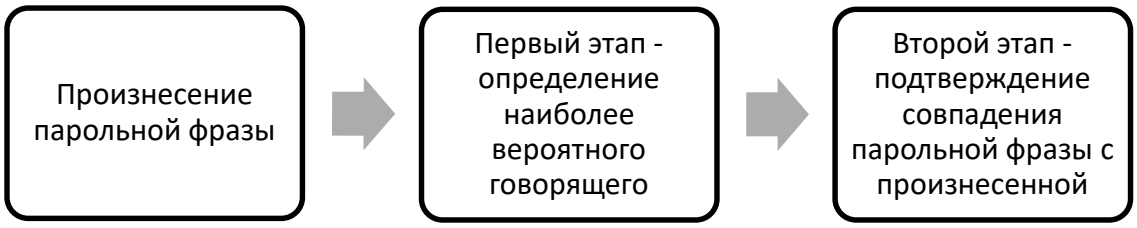

Рисунок 1. – Общая схема работы системы

1. Первый этап – идентификация. Первый пункт в решении задачи идентификации – создание базы записей голоса для каждого человека, который впоследствии будет подвергаться идентификации. Для унификации голосовых записей для обоих этапов идентификации было решено использовать цифры в устном формате. Предполагается, что этого будет достаточно для достижения приемлемой точности при тестировании эффективности данного подхода. Здесь, при условии, что запись проводится на непрофессиональном с точки зрения качества звука оборудовании, сразу же появляются несколько проблем: искажения, связанные с шумами или фоновыми звуками, выделение

полезной части записи и т.д. Большинство вопросов можно решить, используя, например, готовый датасет [2], но для более строгого тестирования системы обучение и тестирование итоговой модели проведено в том числе и в условиях низкокачественных записей.

Итак, задача состоит в построении классификатора, отображающего пространство неких признаков записи голоса в пространство действительных чисел, описывающее классы – т.е. голоса записанных людей.

На рис. 2 представлен процесс получения данных для классификации голоса.

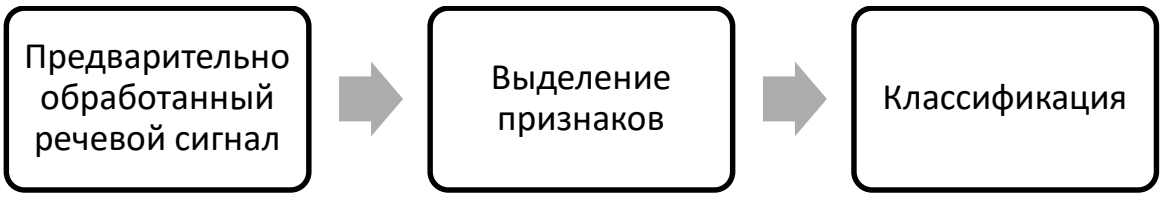

Рисунок 2. – Схема работы классификатора

Предобработка исходных данных и состоит в вейвлет-фильтрации и подавлении выбросов на записи. Длительность записей в идеале должна быть как можно большей, кроме того, длительности отдельных записей могут не совпадать, поэтому использование значений самого речевого потока в качестве вектора признаков необоснованно усложнит задачу. В качестве вектора признаков были выбраны мел-частотные кепстральные коэффициенты. [1]. Количество вычисляемых коэффициентов может быть ограничено любым значением, благодаря чему можно сжать вектор признаков до небольшого размера. Полученные векторы можно усреднить, либо использовать как поток входных данных для рекуррентной нейронной сети.

Следующий шаг – создание классификатора. Имея вычисленные векторы коэффициентов для занесенных в базу записей голосов, необходимо создать модель, способную отнести поступающий на вход системы речевой поток к одному из известных ей голосов. В данной работе обучающая выборка состояла из 1700 записей цифр от 0 до 9, тестовая – из 430 записей. Помимо взятых из бесплатных библиотек записей, в набор входили 360, сделанных самостоятельно на относительно низкокачественной аппаратуре. Всего для обучения системы был использован 21 голос. Также были использованы еще 3 неизвестных системе голоса для тестирования. Оценка точности работы проводилась методом кросс-валидации на одиночных записях, на наборе из 6 цифр, произнесенных одним голосом двумя способами:

1) Предсказанием результата для каждой отдельной записи из набора и усреднением полученных результатов.

2) Предсказанием результата для записи, полученной слиянием всего набора и усреднением полученных признаков

В качестве классификаторов были выбраны методы KNN – k ближайших соседей, SVM – опорных векторов, MLP – многослойный перцептрон.

В таблице 1 представлены результаты тестирования вышеуказанных классификаторов.

С каждым из вышеуказанных способов были проведены два варианта испытаний – с порогом принятия решения и без. В первом случае приходится жертвовать точностью распознавания известных голосов в угоду возможности уже на первом этапе запрещать

доступ в систему при не прошедшем порог входном сигнале. Таким образом уменьшается общее количество ложноположительных срабатываний за счет увеличения числа ложноотрицательных. Во втором случае же ошибок второго рода не может быть по определению, т.к. решение модели принимается как верное. Соответственно, при таком подходе уменьшается количество ложноотрицательных срабатываний за счет увеличения числа ложноположительных. При этом на неизвестных системе голосах будет 100% ошибок первого рода (в табл.1 не указано).

Таблица 1. – Результаты тестирования

Метод оценки/Модель	KNN	SVM	MLP
Средняя точность 10-Fold кросс-валидации (только известные голоса)	95%	96%	97%
Средняя точность на наборах из 6 записей при усреднении предсказаний модели без порога (только известные голоса)	96%	98%	98%
Средняя точность на наборах из 6 записей при усреднении предсказаний модели с порогом [доля ошибок 1 рода на неизвестных голосах]	84% [54%]	84% [58%]	85% [53%]
Средняя точность на наборах из 6 записей при усреднении вектора признаков без порога (только известные голоса)	92%	92%	95%
Средняя точность на наборах из 6 записей при усреднении вектора признаков с порогом [доля ошибок 1 рода на неизвестных голосах]	85% [44%]	86% [73%]	90% [56%]

Примечание. Под точностью здесь подразумевается отношение числа верно распознанных голосовых записей (наборов записей) к общему числу записей (наборов записей), на которых было проведено тестирование. Для тестирования с порогом в квадратных скобках указано отношение прошедших порог наборов записей ко всем тестируемым наборам при тестировании ТОЛЬКО на неизвестных голосах – т.е. только процент ошибочных срабатываний системы (при отсутствующем пороге отношение будет один к одному – 100%).

2. Второй уровень – верификация. Итак, модель способна практически точно определить. Остались только остаются ошибки первого рода, уменьшения количества которых можно добиться с помощью необходимости произнесения некой случайной парольной фразы, которая, опять же, для простоты будет набором цифр.

Входные данные для обучения второй модели, которая должна на элементарном уровне понимать, какую цифру ей сказали, или сказали что-то не похожее на цифру, будут очень похожи на данные из предыдущего пункта, но с бóльшим разрешением по времени. Дальнейший выбор обычно лежит между марковскими моделями и рекуррентными нейронными сетями, на которых в большинстве случаев построены современные системы распознавания слитной речи.

Тем не менее, в данной работе для реализуемой системы возможность модели понимать, какую цифру ей говорят, должна быть максимально зависимой голоса от говорящего, т.е. обладать наименьшей обобщающей способностью. Таким образом, в качестве классификатора выбран многослойный перцептрон.

При достаточной длительности парольной фразы (6-7 и более цифр) можно полностью избавиться от ошибок первого рода.

Таблица 2. – результаты тестирования второй модели

Метод оценки/Модель	MLP
Средняя точность на записях голоса, на котором обучалась модель	92%
Средняя точность на голосах, не входящих в обучающую выборку	16%

Таблица 3. – результаты тестирования системы целиком

Средняя точность распознавания на всех имеющихся голосах, в которые входили как голоса из обучающей выборки, так и неизвестные системе [доля ошибок 1/2 рода]	88% [0% / 12%]

Итоговые результаты получаются несколько хуже ожидаемых, т.к. для оценки были выбраны фиксированные тестовая и обучающая выборки. Однако при длине парольной фразы в 7 цифр и пороге в 6 верно распознанных количество ошибок первого рода составляет 0, соответственно получившаяся система является безопасной с точки зрения ложного предоставления доступа, а итоговый результат можно улучшить, применяя более сложные модели распознавания речи, не рассмотренные здесь.

Таким образом, если принять предположение о том, что исходная выборка записей голосов репрезентативна, в качестве верного, модель, осуществляющая верификацию, будет показывать схожие результаты на большем множестве голосов, вероятность верного распознавания может достигать точности распознавания не хуже 88 %, согласно проведенным в работе исследованиям.

Литература

1. Методы распознавания речи – Режим доступа: https://moluch.ru/archive/130/36213/.– Дата доступа: 27.02.2018.
2. Free spoken digit dataset – Режим доступа: https://github.com/Jakobovski/free-spoken-digit-dataset. – Дата доступа: 23.02.2018.
3. The future of voice search – Режим доступа: https://searchenginewatch.com/2017/04/06/ubiquitous-and-seamless-the-future-of-voice-search/. –Дата доступа: 27.02.2018.