УДК 004.93

МОДИФИКАЦИЯ VGG-АРХИТЕКТУРЫ В ЗАДАЧАХ УНИМОДАЛЬНОЙ И МУЛЬТИМОДАЛЬНОЙ БИОМЕТРИИ

Стефаниди А.Ф., *аспирант Ярославского государственного университета им. П.Г. Демидова, e-mail: antonstefanidi@mail.ru;*
Приоров А.Л., *д.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: andcat@yandex.ru;*
Топников А.И., *к.т.н., Ярославский государственный университет им. П.Г. Демидова, e-mail: topartgroup@gmail.com;*
Хрящев В.В., *к.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: vhr@yandex.ru.*

MODIFICATION OF VGG-ARCHITECTURE FOR UNIMODAL AND MULTIMODAL BIOMETRICS

Stefanidi A. F., Priorov A. L., Topnikov A. I., Khryashchev V. V.

The paper considers a problem of personality recognition using neural network approaches. We developed two algorithms based on the analysis of audio and video data. Both approaches are implemented by using a modification of the VGG convolutional neural network. The first identification method is unimodal. The convolutional neural network CNN-VGGS analyzes the mel-frequency cepstral coefficients of the speech signal. The second algorithm is a multimodal approach based on the original architecture of the bidirectional neural network CNN-VGGMulti. The method classifies a person by merging voice and face analysis. The research results proved the effectiveness of the multimodal algorithm in the problem of personality recognition and this method can be used in real biometric systems.

Key words: digital speech processing, digital image processing, speaker identification, face recognition, convolutional neural network, unimodal biometrics, bimodal biometrics.

Ключевые слова: цифровая обработка речевых сигналов, цифровая обработка изображений, идентификация диктора, распознавание лиц, сверточная нейронная сеть, унимодальная биометрия, бимодальная биометрия.

Рассматривается задача распознавания личности с применением нейросетевых подходов. Предложены два алгоритма на основе анализа аудио- и видеоданных. Оба подхода реализованы с использованием модификации сверточной нейронной сети VGG-архитектуры. Первый метод идентификации личности является унимодальным. Сверточная нейронная сеть CNN-VGGS анализирует мел-частотные кепстральные коэффициенты речевого сигнала. Второй алгоритм представляет собой мультимодальное решение на базе оригинальной архитектуры двунаправленной нейронной сети CNN-VGGMulti. Метод классифицирует личность, комбинируя результаты анализа голоса и лица. Результаты исследования доказали эффективность мультимодального алгоритма в задаче распознавания личности. Решение может быть использовано для разработки реальных биометрических систем.

Введение

В настоящее время существует большое количество сервисов, приложений и услуг, использующих методы биометрической идентификации. Связано это с тем, что данная технология позволяет достаточно точно аутентифицировать пользователя и защитить его персональные данные от несанкционированного доступа. Как правило, это методы анализа по отпечатку пальцев, изображению лица или голосу. Однако любая унимодальная система имеет свойственный ей ряд ограничений. Биометрия на основе анализа отпечатков пальцев является контактным методом, что уменьшает область практической применимости данной технологии. Системы распознавания пользователя по изображению лица имеют сильную зависимость от уровня освещенности, ракурса, качества фоторегистратора, они также чувствительны к возрастным изменениям и мимике. Система идентификации диктора зависит от эффектов канала передачи информации и микрофона, физиологических особенностей говорящего, акустических свойств окружающей среды [1-7].

В работе описывается разработка метода мультимодальной идентификации личности с использованием лицевой и голосовой биометрии. Такой подход позволяет создать систему бесконтактного сбора биометрических данных, обладающую высоким уровнем надежности. Использование двух биометрических параметров существенно уменьшает вероятность фальсификации данных [4-6]. Первая часть данной работы посвящена рассмотрению унимодального метода идентификации личности с использованием речевых сигналов. Во второй части рассматривается бимодальный алгоритм, основанный на объединении лицевой и голосовой биометрии.

Нейросетевой подход стал одним из главных инструментов в решении задач детектирования, распознавания и сегментации объектов. В частности, методы и алгоритмы на основе нейронных сетей показывают высокие результаты идентификации людей с использованием цифровых изображений и речевых сигналов [8-13].

Рис. 1. Примеры изображений лиц из набора данных VoxCeleb1

Такие сети также используются в задачах обработки и анализа текстов, в медицине, биохимических исследованиях, робототехнике. Практическую значимость нейросетевых подходов сложно переоценить. Данное исследование также основывается на применении сверточных нейронных сетей.

Целью работы является разработка алгоритмов унимодальной и мультимодальной биометрии на основе анализа речевых сигналов и цифровых изображений лиц.

Описание набора данных

Для проведения эксперимента использовалась популярная база VoxCeleb1 – аудиовизуальный набор данных, состоящий из коротких фрагментов человеческой речи и цифровых изображений лиц, извлеченных из видеоинтервью [14]. Процесс подготовки данной базы представляет собой оригинальный метод сбора биометрических данных и состоит из нескольких этапов. Вначале авторы выбрали список известных людей из базы лиц VGG Face, который насчитывает 2622 личности. Вторым этапом для каждого класса выгружались по 50 самых популярных YouTube-видео. Для того, чтобы поиск был более точным, авторы комбинировали имя знаменитости со словом «интервью» при каждом поисковом запросе. Это позволило повысить вероятность того, что конкретный человек действительно присутствует на видео.

Детектирование и выравнивание лиц выполнялись с использованием HOG-детектора и ансамбля регрессионных деревьев [15, 16]. Для трекинга лиц применялись подходы, описанные в [17, 18]. Следующим этапом из видео выделялся аудиосигнал, содержащий речь конкретного человека. Идея заключается в синхронизации движения рта на видео и в речи. Тем самым можно сказать, какому именному лицу принадлежит образец речевой активности. Для этого использовалась нейронная сеть SyncNet [19]. Такой подход способен фильтровать дубляж и закадровый голос. На последнем этапе выполнялась верификация детектированных лиц. В качестве классификатора применялась сеть CNN VGG Face, обученная на наборе данных VGG Face [11].

Тестовая база VoxCeleb1 содержит записи речи спикеров, охватывающих широкий спектр различных национальностей, стиля произношения, профессий и возраста. Набор включает более 150000 речевых сигналов для 1251 класса. Помимо речи база VoxCeleb1 содержит набор цифровых изображений лиц, которые были детектированы и вырезаны в процессе обработки видеороликов (рис. 1). Суммарное количество изображений составляет более 1,2 млн. Изображения обладают следующими свойствами: лица имеют разный угол поворота/наклона головы, цвет лица/волос, наличие/отсутствие очков/бороды и усов; съемка с различными сценами и степенью освещенности [14].

Благодаря наличию хорошо структурированной и размеченной базы данных VoxCeleb1, состоящей из цифровых изображений лиц и речевых сигналов, открывается возможность разработки мультимодальной системы идентификации на основе двух биометрических параметров – лица и голоса. Поскольку набор является достаточно крупным и требовательным к вычислительным ресурсам, количество классов было уменьшено с 1251 до 50.

Предобработка данных и описание архитектур сетей

В табл. 1 указан объем исследуемой части аудиовизуальной базы VoxCeleb1. Речевые сигналы представлены в формате wav с частотой дискретизации 16 кГц и уровнем квантования по амплитуде 16 бит. Для того чтобы преобразованные данные можно было компоновать в массивы одной размерности, все речевые сигналы должны быть одной длительности. Из каждого речевого сигнала выделялся фрагмент длительностью в 3 секунды согласно методике, изложенной в [14]. Фрагменты выбирались по принципу случайного выделения из оригинального сигнала.

Таблица 1. Статистика анализируемой части набора данных VoxCeleb1

	Обучение	Проверка	Тест	Общее
Изображения	43243	2382	2447	48072
Речевые сигналы	5730	271	322	6323

Оригинальные звуковые данные представляют собой изменение амплитуды колебаний во времени, что является не самой информативной формой представления речевого сигнала, поэтому в работе использовались мел-частотные кепстральные коэффициенты (МЧКК). При решении задачи автоматического распознавания речи, как правило, применяется 40-80 фильтров. В работе сформирован банк из 80 треугольных мел-фильтров (рис. 2). В итоге каждый речевой сигнал представлялся матрицей размером 80x301 [20, 21].

Для повышения обобщающей способности обучаемых моделей нейронных сетей применялся метод синтетического увеличения данных. Речевые сигналы подвергались искажениям и преобразованиям: добавление аддитивного белого гауссовского шума; смещение и

растяжение сигнала по времени; изменение высоты тембра; использование эффекта реверберации, позволяющего искусственным образом изменять свойства аудиосигнала, меняя представления о масштабе и глубине акустической сцены; применение медианной фильтрации для разделения гармонических и перкуссионных компонент сигнала (HPSS, Harmonic Percussive Source Separation).

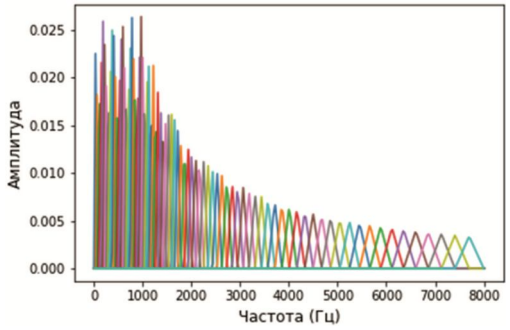

Рис. 2. Банк из 80 треугольных мел-фильтров

Дополнительно, для повышения вариативности данных, использовался открытый набор звуковых сигналов Urban Sound Dataset (UrbanSound8K). Набор включает 8732 записи, каждая длиной менее 4-х секунд, представляющие собой характерные для города звуки: работа кондиционера, автомобильный гудок, игра детей, лаянье собаки, работа двигателя на холостом ходу, выстрел пистолета, работа отбойного молотка, сирена, звуки уличной музыки. Речевые сигналы случайным образом смешивались с сигналами из набора UrbanSound8K [22-25].

В настоящее время публикуется большое количество интересных исследований на основе анализа наборов данных VoxCeleb1 и VoxCeleb2. Высокие результаты при проведении экспериментов получаются с использованием крупных архитектур: ResNet18, ResNet34, ResNet50 [14, 26]. Для решаемой задачи они не подойдут, поскольку исследуется лишь небольшая выборка набора данных VoxCeleb1, которая составляет менее 5 % от общего объема. Если применять очень глубокие архитектуры к такому малому набору данных, очевидно, что сеть будет переобучаться и обладать слабой обобщающей способностью. Поэтому для данного исследования спроектированы более компактные сверточные нейронные сети.

На рис. 3 представлена CNN-VGGS – архитектура сети, используемая для создания унимодальной системы распознавания личности на основе анализа речевого сигнала. Сверточная нейронная сеть является очень компактной, поскольку содержит менее 0,3 млн. весовых коэффициентов. Для сравнения: сеть ResNet18 содержит более 11 млн. весов, а сеть ResNet50 – более 25 млн. весов.

Для решения задачи мультимодальной идентификации личности по двум биометрическим параметрам спроектирована архитектура CNN-VGGMulti (рис. 4).

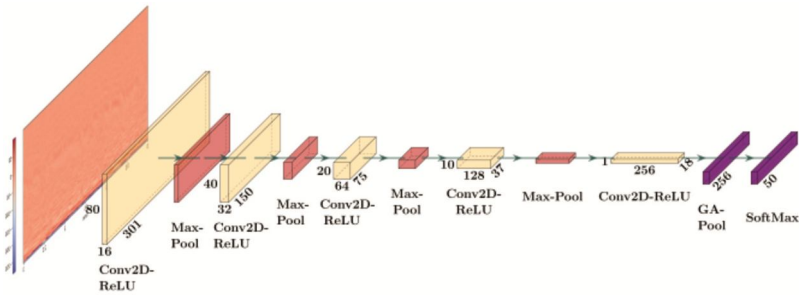

Рис. 3. Архитектура сверточной нейронной сети CNN-VGGS

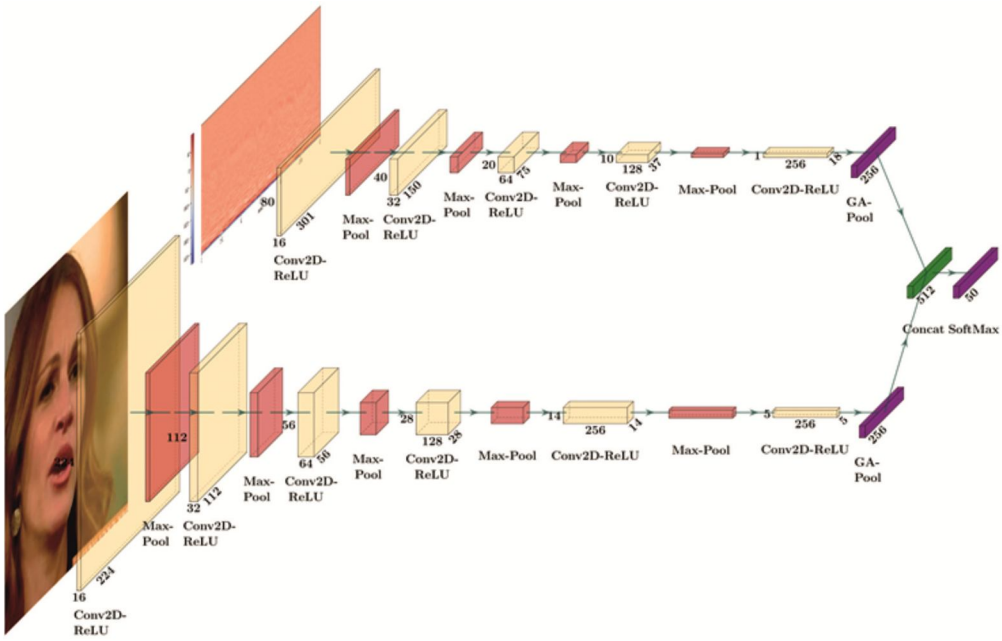

Рис. 4. Архитектура сверточной нейронной сети CNN-VGGMulti

Данная сеть имеет два входа: один для приема цифровых изображений лиц размером 224x224x3, другой – для МЧКК-представления речевых сигналов. Каждый из потоков сети CNN-VGGMulti имеет слой глобального усреднения (GA-Pool, Global Average Pooling). На выходе этих слоев формируются векторы одинаковой размерности по 256 значений, которые далее объединяются с использованием слоя конкатенации Concat в один общий вектор размерностью 512. Это делается для того, чтобы признаки, формируемые разными модальностями, имели одинаковое, равновесное влияние на итоговый результат классификации. Сеть CNN-VGGMulti также является относительно компактной и содержит 0,95 млн. весовых коэффициентов.

Результаты исследования унимодального алгоритма на основе голосовой биометрии

Опишем процесс обучения нейронной сети CNN-VGGS, а также проведем анализ результатов тестирования. В качестве метода оптимизации весовых параметров применялся Adam (Adaptive Moment Estimation) [27]. В процессе обучения устанавливались следующие гиперпараметры: скорость сходимости алгоритма оптимизации 0,001, размер батча 32, количество эпох 100. На рис. 5 представлен процесс обучения сети с использованием метрики оценки качества accuracy. Из результатов видно, что на тренировочном наборе данных доля правильных ответов составляет 99,84 %, однако на проверочном множестве оценка имеет значение в 60,89 %, а на тестовом множестве 51,55 %. Такой результат свидетельствует о переобучении модели и ее низкой обобщающей способности. Также видно, что кривая обучения выходит на плато и увеличение количества эпох не дает существенного улучшения в работе.

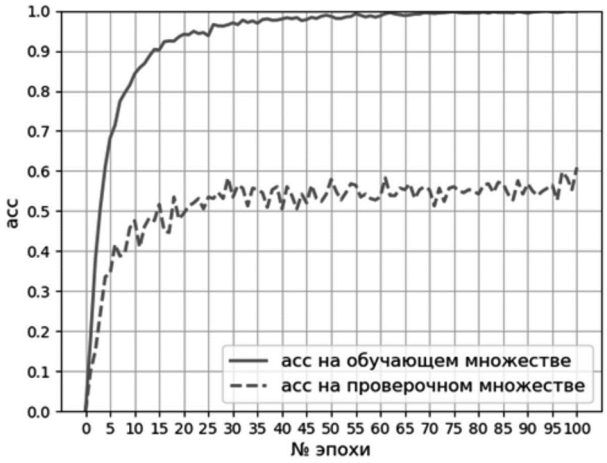

Рис. 5. Кривые изменения доли правильных ответов в процессе обучения сети CNN-VGGS

Одним из распространенных методов борьбы с переобучением является использование одной из разновидностей регуляризации – прореживание слоев. Проведен ряд экспериментов с применением данного подхода, однако получить каких-либо улучшений в точности работы сети не удалось.

Еще одним общепринятым методом повышения обобщающей способности модели является увеличение обучающего набора данных. Новые образцы можно искать в открытых источниках и сети Интернет или синтетически сгенерировать с использованием различных типов преобразований речевого сигнала. В данном исследовании звуковые сигналы подвергались искажениям и изменениям, описанным ранее. На рис. 6 представлен процесс обучения сети на сгенерированных данных. Применение метода искусственного аугментирования позволило немного улучшить оценку на проверочном и тестовом множествах – 64,95 % и 57,76 % соответственно. Однако модель по-прежнему обладает слабой обобщающей способностью и не способна с высокой точностью определить диктора на новых образцах.

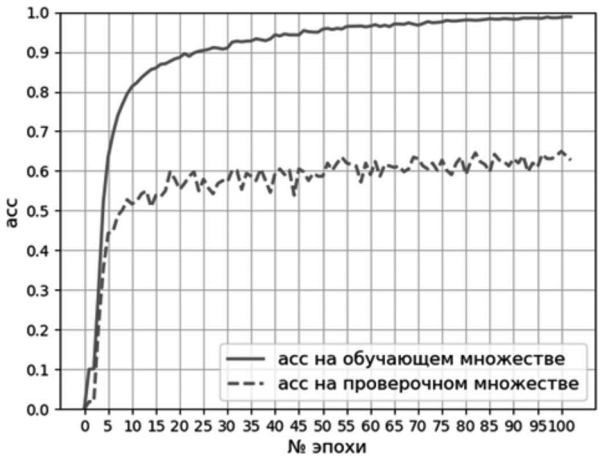

Рис. 6. Кривые изменения доли правильных ответов в процессе обучения сети CNN-VGGS на аугментированных данных

Поскольку метод синтетического увеличения аудиоданных не дал качественного улучшения точности при распознавании личности было решено добавить еще один биометрических параметр, который бы явным образом характеризовал человека. В результате был реализован метод идентификации на основе комбинированного анализа речевых сигналов и цифровых изображений лиц.

Результаты исследования мультимодального алгоритма на основе лицевой и голосовой биометрии

Поскольку унимодальный метод распознавания диктора не продемонстрировал высоких результатов на проверочном и тестовом множествах был разработан мультимодальный алгоритм. Для этого реализована архитектура CNN-VGGMulti, состоящая из двух веток, имеющих один общий выход. Ветка для анализа речевых данных представляет собой сверточную нейронную сеть CNN-VGGS. Ветка для анализа изображений лиц также представляет собой небольшую сверточную нейронную сеть. Каждая из веток формирует 256-мерный вектор на основе соответствующего ей входа с использованием МЧКК-представления речевого сигнала или цифрового изображения лица. Далее эти вектора объединяются в общий вектор размерностью 512. На входе сети CNN-VGGMulti имеются тензоры размером 80x301x1 и 224x22x3. В качестве метода оптимизации весовых параметров, также, как и для унимодального

алгоритма, применялся метод оптимизации весовых параметров Adam. В процессе обучения устанавливались следующие гиперпараметры: скорость сходимости алгоритма оптимизации 0,001, размер батча 8, количество эпох 100. На рис. 7 представлен процесс обучения мудьтимодального алгоритма на основе сверточной нейронной сети CNN-VGGMulti.

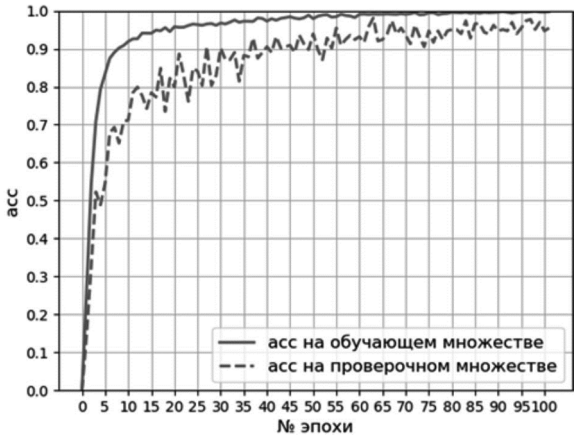

Рис. 7. Кривые изменения доли правильных ответов в процессе обучения сети CNN-VGGMulti

Мультимодальная модель продемонстрировала высокие результаты работы. Оценка на обучающей выборке составила 99,88%, на проверочном и тестовом наборах данных – 98,11% и 97,19% соответственно. Использование комбинированного подхода на основе анализа речевой и лицевой биометрии позволило значительно повысить обобщающую способность классификатора. Обученная модель делает предсказание, основываясь одновременно на двух независимых особенностях человека. Результаты исследования могут быть использованы для проектирования коммерческих систем распознавания личности.

Заключение

В работе рассмотрен вопрос классификации личности с использованием методов унимодальной и бимодальной биометрии. В качестве источника аудиовизуальной информации выбран современный набор данных VoxCeleb1. Для проведения исследования из данной базы выделены 50 уникальных классов. Вся база VoxCeleb1 в исследовании не применялась, что позволило более динамично проводить опытные работы на всех этапах эксперимента. Анализ речевых сигналов основывался на выделении мел-частотных кепстральных коэффициентов.

В ходе работы спроектированы две архитектуры сверточных нейронных сетей. Для унимодальной идентификации личности на основе только речевых сигналов применялась сверточная нейронная сеть CNN-VGGS. Исследование показало, что обученная модель имеет слабую обобщающую способность, показав на тестовом множестве точность 51,55 %. Для борьбы с переобучением применялся метод прореживания слоев и искусственная аугментация данных, однако существенным образом это не повлияло на качество работы. Модель, обученная на новых синтезированных данных, показала на тестовом множестве точность 57,76 %.

Для повышения точности идентификации реализован мультимодальный алгоритм распознавания на основе анализа речевых сигналов и цифровых изображений лиц. Разработана архитектура двунаправленной сверточной нейронной сети CNN-VGGMulti. Мультимодальное решение продемонстрировало высокий уровень точности на проверочном и тестовом наборах данных – 98,11 % и 97,19 % соответственно. Результаты работы бимодального алгоритма свидетельствуют о возможности практической применимости данного подхода в реальных коммерческих системах распознавания личности.

На следующем этапе исследования планируется проведение эксперимента с использованием всего набора данных VoxCeleb1. Для решения задачи распознавания личности будет применен метод комбинирования биометрических параметров, описанный в данной статье, но с использованием уже более глубоких топологий сетей ResNet18 и ResNet34. Дополнительно будут представлены новые методы объединения биометрических параметров.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-37-90158.

Литература

1. Хайкин С. Нейронные сети: Полный курс. Пер. с англ. Н.Н. Куссуль, А.Ю. Шелестова. 2-е изд. – М.: Издательский дом Вильямс, 2008. 1103 с.

2. Матвеев Ю.Н. Технология биометрической идентификации личности по голосу и другим модальностям // Вестник МГТУ им. Н.Э. Баумана. Сер. «Приборостроение». 2012. № 3. С. 46-61.

3. Козлов А.В., Кудашев О.Ю., Матвеев Ю.Н., Пеховский Т.С. Система идентификации дикторов по голосу для конкурса NIST SRE 2013 // Труды СПИИРАН, 2013. № 2. С. 350-370.

4. Хрящев В.В., Приоров А.Л., Стефаниди А.Ф., Топников А.И. Разработка и исследование алгоритмов обработки и распознавания речевых сигналов и изображений для систем мультимодальной биометрии // Цифровая обработка сигналов. 2017. № 3. С. 45-49.

5. Khryashchev V., Topnikov A., Stefanidi A., Priorov A. Bimodal person identification using voice data and face images. In Proceedings SPIE 11041, Eleventh International Conference on Machine Vision, 2018, Web: https://doi.org/10.1117/12.2523138.

6. Stefanidi A., Topnikov A., Tupitsin G., Priorov A. Application of convolutional neural networks for multimodal identification task, 26th Conference of Open Innovations Association FRUCT, 423-428, (2020); doi: 10.23919/FRUCT 48808.2020.9087458

7. Tupitsin G., Topnikov A., Priorov A. Two-step noise reduction based on soft mask for robust speaker identification, In Proceedings 18th Conference of Open Innovations Association FRUCT. 2016. pp. 351-356.

8. Ault S.V., Perez R.J., Kimble C.A., Wang J. On Speech Recognition Algorithms. International Journal of Machine Learning and Computing, vol. 8, no. 6, 2018, pp. 518-523.

9. Bunrit S., Inkian T., Kerdprasop N. Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network. International Journal of Machine Learning and Computing, Vol. 9, no. 2, 2019, pp. 143-148.

10. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang S., Karpathy A., Khosla A., Bernstein M., Berg A.C., Li F.F. Imagenet large scale visual recognition challenge. IJCV, 2015.

11. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, 2015, Web: https://arxiv.org/abs/ 1409.1556v6.

12. Parkhi O.M., Vedaldi A., Zisserman A. Deep face recognition. In Proceedings British Machine Vision Conference, 1, 41.1-41.12 10.5244, 2015, pp. 29-41.

13. Sun Y., Ding L., Wang X., Tang X. DeepID3: Face recognition with very deep neural networks, 2015, Web: https://arxiv.org/ abs/1502.00873.

14. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset, 2017, Web: https://arxiv.org/abs/ 1706.08612v2.

15. King D.E. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, vol. 10, pp. 1755-1758, 2009.

16. Kazemi V., Sullivan J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867-1874, 2014.

17. Chung J. S., Zisserman A. Lip reading in the wild. In Proceedings of the Asian Conference on Computer Vision, 2016.

18. Everingham M., Sivic J., Zisserman A. Taking the bite out of automatic naming of characters in TV video. Image and Vision Computing, vol. 27, no. 5, 2009.

19. Chung J. S., Zisserman A. Out of time: automated lip sync in the wild. In Workshop on multi-view lip-reading, ACCV, 2016.

20. Logan B. Mel frequency cepstral coefficients for music modeling. In International Symposium Music Information Retrieval, 2000.

21. Kopparapu S.K., Laxminarayana M. Choice of Mel filter bank in computing MFCC of a resampled speech. 2010. pp. 121-124, doi: 10.1109/ISSPA.2010.5605491.

22. Salamon J., Bello J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. In IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.

23. Park D.S., Chan W., Zhang Y., Chiu C.C., Zoph B., Cubuk E.D., Le Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. 2019.

24. Ko T., Peddinti V., Povey D., Khudanpur S. Audio augmentation for speech recognition. In INTERSPEECH-2015, 2015. pp. 3586-3589.

25. Salamon J., Jacoby C., Bello J.P. A Dataset and Taxonomy for Urban Sound Research. 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.

26. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition. In Proceedings Interspeech, 2018, pp. 1086-1090.

27. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization. 2017, Web: https://arxiv.org/ abs/1412. 6980v9.

НОВЫЕ КНИГИ

Подкорытов А.Н.

Методы местоопределения потребителя в глобальных навигационных спутниковых системах / Учебное пособие для вузов – М.: Изд-во «Горячая линия-Телеком», 2020 г. – 136 стр.: ил.

Приведена классификация основных методов местоопределения в глобальных навигационных спутниковых системах. Рассмотрены абсолютные и относительные методы местоопределения различной точности и оперативности по измерениям систем ГЛОНАСС и GPS, в том числе с использованием систем широкозонной дифференциальной коррекции. Описано местоопределение как статического, так и динамического потребителя. Значительное внимание уделено высокоточному абсолютному местоопределению. Изложены теоретические и методические материалы к лабораторным работам по местоопределению потребителя в глобальных навигационных спутниковых системах.

Для студентов вузов, обучающихся по укрупненным группам специальностей и направлениям подготовки: 24.00.00 – «Авиационная и ракетно-космическая техника», 11.00.00 – «Электроника, радиотехника и системы связи». Будет полезно аспирантам и специалистам соответствующих направлений и научной специальности ВАК 05.12.14.

Уважаемые коллеги!

Для тех, кто не успел оформить подписку на второе полугодие 2020 года через АО «Роспечать», сохраняется возможность приобретения журналов непосредственно в редакции по адресу: г. Москва, ул. Авиамоторная, дом 8, Научный Центр МТУСИ, ком. 612. Российское научно-техническое общество радиотехники, электроники и связи им. А.С. Попова, метро «Авиамоторная», или оформить Заказ в соответствии с требованиями, выставленными на сайте журнала: www.dspa.ru.

Справки по телефону: (+7 903) 201-53-33 (Самсонов Геннадий Андреевич). E-mail: rntores@mail.ru