Автоматическая верификация диктора по произвольной фразе с применением свёрточных глубоких сетей доверия

И.А. Рахманенко [1], А.А. Шелупанов [1], Е.Ю. Костюченко [1]
[1] Томский государственный университет систем управления и радиоэлектроники, 634050, Россия, Томская область, г. Томск, пр. Ленина, д. 40

Аннотация

Данная статья посвящена применению свёрточных глубоких сетей доверия в качестве средства извлечения речевых признаков из аудиозаписей для решения задачи автоматической, текстонезависимой верификации диктора. В работе описаны область применения и проблемы систем автоматической верификации диктора. Рассмотрены типы современных систем верификации диктора, основные типы речевых признаков, используемых в системах верификации диктора. Описана структура свёрточных глубоких сетей доверия, алгоритм обучения данной сети. Предложено применение речевых признаков, извлекаемых из трёх слоёв обученной свёрточной глубокой сети доверия. Данный подход основан на применении методов анализа изображений как к уже выделенным признакам речевого сигнала, так и для их выделения из слоёв нейронной сети. Произведены экспериментальные исследования предложенных признаков на двух речевых корпусах: собственном речевом корпусе, включающем аудиозаписи 50 дикторов, и речевом корпусе TIMIT, включающем аудиозаписи 630 дикторов. Была произведена оценка точности предложенных признаков с применением классификаторов различного типа. Непосредственное применение данных признаков не дало увеличения точности по сравнению с использованием традиционных речевых признаков, таких как мел-кепстральные коэффициенты. Однако применение данных признаков в составе ансамбля классификаторов позволило достичь уменьшения равной ошибки 1-го и 2-го рода до 0,21 % на собственном речевом корпусе и до 0,23 % на речевом корпусе TIMIT.

Ключевые слова: распознавание диктора, верификация диктора, Гауссовы смеси, GMM-UBM-система, речевые признаки, обработка речи, глубокое обучение, нейронные сети, распознавание образов.

Цитирование: **Рахманенко, И.А.** Автоматическая верификация диктора по произвольной фразе с применением свёрточных глубоких сетей доверия / И.А. Рахманенко, А.А. Шелупанов, Е.Ю. Костюченко // Компьютерная оптика. – 2020. – Т. 44, № 4. – С. 596-605. – DOI: 10.18287/2412-6179-CO-621.

Citation: Rakhmanenko IA, Shelupanov AA, Kostyuchenko EYu. Automatic text-independent speaker verification using convolutional deep belief network. Computer Optics 2020; 44(4): 596-605. DOI: 10.18287/2412-6179-CO-621.

Введение

Одной из сложных и нерешённых задач в области цифровой обработки аудиосигналов является задача автоматической верификации и идентификации диктора. Существуют различные ограничения в применении современных систем верификации диктора, к которым можно отнести ограничения по акустическим условиям, в которых обучаются и используются модели диктора, ограничения по состоянию самого диктора, соотношению сигнал-шум и др. Все эти ограничения главным образом влияют на основную характеристику системы верификации диктора – точность.

Требования к точности систем верификации в основном зависят от области применения таких систем. В последнее время проявляется тенденция расширения требований к количеству пользователей системы верификации, что связано с их применением в таких областях, как дистанционное банковское обслуживание, системы контроля доступа на крупных предприятиях, системы биометрической многофакторной аутентификации. С каждым годом требования к точности верификации повышаются, требуя исследований и разработки новых методов верификации диктора. Таким образом, для удовлетворения нужд потребителей данных систем и выдвигаются высокие требования по точности верификации диктора.

Задача распознавания диктора включает в себя две подзадачи: идентификацию и верификацию. Автоматическая верификация диктора – это подтверждение личности по голосу в соответствии с предъявленным им идентификатором (обычно, именем данного диктора). Отличие же автоматической идентификации диктора заключается в том, что изначально неизвестен идентификатор диктора, соответственно, система должна сама определить, кем является данный диктор – законным пользователем, зарегистрированным в системе, или нарушителем (в случае решения

задачи открытой идентификации) [1]. Система автоматической текстонезависимой верификации диктора, представленная в данной работе, решает задачу верификации закрытого множества дикторов, определяя, присутствует ли на аудиозаписи голос заявленного диктора или нет. В данном случае существование дикторов, не зарегистрированных в системе, не принимается во внимание.

В данной работе ставится задача исследования применимости новых признаков, извлекаемых из аудиозаписей с помощью свёрточных глубоких сетей доверия (*Convolutional Deep Belief Network*, СГСД), с целью автоматической текстонезависимой верификации диктора по произвольной фразе. Ранее данный вид глубоких нейронных сетей был успешно использован для извлечения признаков из изображений [2, 3], аудио-сигнала [4], электроэнцефалограммы [5].

1. Обзор современных систем верификации диктора

Можно выделить несколько основных частей, из которых состоят системы верификации диктора. К ним относят подсистемы предобработки аудио, извлечения из аудиозаписей признаков, отражающих индивидуальные характеристики голоса диктора, создания моделей дикторов и принятия решений, необходимых для проведения оценки соответствия аудиозаписей соответствующим моделям диктора.

К современным системам верификации диктора по произвольной фразе, которые наиболее часто встречаются в научных работах, можно отнести несколько видов систем: системы, основанные на Гауссовых смесях [6, 7], системы, основанные на применении i-векторов [8 – 12], и системы с применением глубоких нейронных сетей [9 – 11, 13, 14]. Рассмотрим данные виды систем с учётом тех моделей, которые используются при решении задачи верификации диктора по произвольной фразе.

1.1. GMM – UBM - системы

Одной из базовых моделей, используемых при сравнении новых методов верификации диктора с аналогами, являются Гауссовы смеси (ГС, GMM). К особенностям таких моделей относят низкую вычислительную сложность и низкую устойчивость к смене акустических условий. В данной работе одной из используемых моделей являются ГС.

Гауссова смесь – это параметрическая функция плотности вероятности, представленная как взвешенная сумма отдельных Гауссовых плотностей [15]. ГС, состоящая из C плотностей вероятности, может быть представлена формулой:

$$p(x \mid \lambda) = \sum_{i=1}^{C} w_i g(x \mid \mu_i, \Sigma_i), \quad (1)$$

где x – это D-мерный непрерывный вектор данных (признаков), w_i, $i = 1,\ldots,C$ – это вес i-го компонента смеси, $g(x \mid \mu_i, \sum_i)$, $i = 1, ..., C$ – это Гауссова плотность вероятности i-го компонента смеси с вектором математических ожиданий μ_i и ковариационной матрицей $\sum_i$.

Таким образом, полную ГС можно описать множеством векторов математического ожидания, ковариационных матриц и весов смесей каждого компонента модели. ГС можно представить уравнением (2):

$$\lambda = \{w, \mu, \Sigma\}. \quad (2)$$

При решении задачи верификации диктора каждый из дикторов представлен в системе собственной ГС λ.

Универсальная фоновая модель (УФМ, UBM) – это ГС, обученная на большом наборе речевого материала, взятого от большого множества дикторов, ожидаемых системой во время распознавания. Благодаря этому можно использовать УФМ для проверки альтернативной гипотезы, т.е. того случая, когда на записи отсутствует голос заданного диктора. Как и в [15], параметры для УФМ были обучены с помощью EM-алгоритма (*Expectation-Maximization* алгоритм), а для обучения моделей дикторов была использована форма Байесовой адаптации (MAP-адаптация).

1.2. Системы с применением i-векторов

Подход, использующийся в системах верификации с применением i-векторов, заключается в определении единого пространства, отражающего в себе индивидуальные характеристики и голоса диктора и окружения, в котором был записан голос. Данный метод позволяет отследить все изменения, происходящие во время адаптации математических ожиданий УФМ для заданной последовательности окон рассматриваемого отрезка речи. Эта информация моделируется в пространстве малой размерности, называемом пространством полной изменчивости.

Таким образом, в данном методе каждая произнесённая диктором фраза имеет соответствующий вектор M, заданный следующим образом (3):

$$M = m + Tw, \quad (3)$$

где m – дикторо- и каналонезависимый супервектор (например, универсальная фоновая модель); T – квадратная матрица малого порядка и w – случайный вектор с нормальным распределением $N(0, I)$. Компоненты вектора w являются полными факторами, а сам вектор называется вектором идентичности или i-вектором (*i-vector*). Данный вектор является скрытой переменной, которая может быть задана апостериорным распределением с использованием статистики Баума–Велша.

Компоненты i-вектора отражают изменения в компонентах Гауссовой смеси универсальной фоновой модели (супервектор m), произошедшие после адаптации УФМ к заданной фразе диктора. При про-

ведении верификации диктора *i*-векторы являются соответствующими отрезку речи диктора признаками. После извлечения данных векторов они подаются на вход классификатору.

Используются несколько видов систем, основанных на данном представлении индивидуальных характеристик голоса диктора. Для вычисления оценки соответствия *i*-векторов применяют вероятностный линейный дискриминантный анализ (PLDA) [8, 9, 13, 14], машины опорных векторов с применением косинусного ядра, косинусное расстояние между *i*-векторами [16] и др.

1.3. Системы с применением глубоких нейронных сетей

В последнее время в задаче распознавания образов активно развиваются методы глубокого обучения, что повлекло за собой формирование тренда на применение глубоких нейронных сетей (ГНС) в системах верификации диктора по голосу. Глубокие нейронные сети используются как для извлечения *i*-векторов [11], так и для извлечения новых признаков, которые формируются нейронной сетью в скрытом слое с меньшим количеством нейронов (*Bottleneck Features*, BNF) [13, 14]. Возможно применение ГНС в качестве отдельного классификатора, обученного с целью идентификации диктора [17]. Кроме того, возможно применение ГНС, обученных для распознавания речи, а затем использующихся для извлечения как BNF, так и признаков, полученных из выходного слоя ГНС [9, 13, 14].

Обычно для данных целей применяют нейронные сети прямого распространения, которые намного больше (более тысячи нейронов в скрытом слое) и намного глубже (5 – 7 скрытых слоёв) традиционных нейронных сетей. Для обучения ГНС применяют алгоритм обратного распространения ошибки и метод стохастического градиентного спуска.

Одним из последних направлений является разработка ГНС, позволяющих решать задачу автоматической верификации диктора без применения дополнительных методов извлечения признаков или методов классификации [18, 19]. Данный подход использует сложные архитектуры ГНС, для которых на вход подаются необработанные аудиосигналы, а их выходом является решение о наличии голоса диктора в аудиозаписи.

2. Извлечение признаков

В настоящее время наиболее актуальными являются методы извлечения признаков из аудиозаписей речи с помощью глубоких нейронных сетей. Часть из таких решений используется только для извлечения признаков, часть является «end-to-end» решениями – как извлекающими признаки, так и принимающими финальное решение [18, 19].

В данной работе для сравнения результатов были использованы как традиционные признаки, включающие мел-кепстральные коэффициенты, так и признаки, извлекаемые с помощью СГСД. Кроме этого, были использованы признаки, рассмотренные в [20], показавшие хорошие результаты в задаче текстонезависимой верификации диктора. В этот вектор входит 28 признаков, включающих мел-кепстральные коэффициенты, их дельты и двойные дельты, коэффициент линейного предсказания, линейные спектральные пары и вероятность вокализации (*voicing probability*). Данный набор признаков был сформирован с помощью жадного метода добавления-удаления признаков [21].

К традиционным признакам, использующимся в различных задачах обработки аудио, и в том числе для автоматической верификации диктора, относят такие признаки, как мел-кепстральные коэффициенты, пары линейного спектра (*line spectral pair*, LSP), кепстральные коэффициенты перцептивного линейного предсказания (*perceptual linear prediction cepstral coefficients* – PLP), кратковременная энергия, формантные частоты, частота основного тона, джиттер, шиммер и др.

Один из самых часто используемых признаков, используемых в научных работах, связанных с обработкой речи и распознаванием диктора, являются мел-кепстральные коэффициенты (МКК, *Mel frequency cepstral coefficients*). Метод мел-частотного кепстрального преобразования спектра был впервые представлен в работе [21]. Наиболее часто используют от 12 до 20 МКК. Кроме того, часто используются дельта и двойные дельта-коэффициенты, которые отражают изменения в мел-кепстральных коэффициентах во времени.

Для вычисления МКК используется следующий процесс: на первом шаге производится разделение аудиозаписи на окна – маленькие части речевого аудиосигнала. Данные окна обрабатываются по отдельности, обработка всего сигнала целиком не производится. Длина такого окна составляет 20 мс, а смещение, по которому сигнал разбивается на окна, составляет 10 мс. После этого производится предобработка сигнала (фильтр верхних частот) и умножение на оконную функцию Хэмминга.

Далее производится дискретное преобразование Фурье (ДПФ) и переход к шкале мел. Частоты f, полученные после ДПФ, переводят к шкале мел f_{mel} с помощью преобразования:

$$f_{mel} = 1125 \ln\left(1 + \frac{f}{700}\right). \quad (4)$$

Преобразование между частотами в герцах и в мелах является линейным до частоты 1000 Гц и логарифмическим выше данной частоты [22]. Для выполнения данного преобразования создается набор треугольных фильтров и вычисляется логарифм энергии

в каждой полосе частот данных фильтров [21]. Последним шагом извлечения МКК является выполнение обратного ДПФ.

В данной работе для сравнения точности системы верификации с применением данных речевых признаков был использован вектор из 13 или 14 мелкепстральных коэффициентов, а также их дельта и двойные дельта-коэффициенты. Вычисление спектральных речевых признаков в данной работе проводилось с помощью библиотеки openSMILE [23].

Однако, по мнению авторов, данные признаки недостаточно полно отражают информацию об индивидуальных характеристиках голоса. Мы считаем, что существуют другие признаки, которые могут содержать дополнительную информацию о дикторе, применение которой может улучшить точность распознавания диктора. В качестве речевых признаков, дополняющих традиционно используемые, мы предлагаем рассмотреть признаки, извлекаемые из аудиозаписи с помощью свёрточной глубокой сети доверия.

3. Свёрточные глубокие сети доверия

Основным отличием СГСД [3, 4] от обычной глубокой сети доверия [24] является применение в качестве слоёв сети свёрточной ограниченной машины Больцмана (CRBM – *Convolutional Restricted Boltzmann Machine*, СОМБ) [3]. Данная нейронная сеть (рис. 1) представляет собой детектор признаков, состоящий из трёх слоёв – видимого слоя V, слоя детекции H и слоя агрегирования P.

Применение дополнительного слоя агрегирования (*pooling*) позволяет уменьшить детализацию подаваемых на следующий скрытый слой данных, что позволяет выделять более крупные особенности в признаках. Это также позволяет уменьшить вычислительную нагрузку на последующих слоях и отфильтровывать случайные шумы.

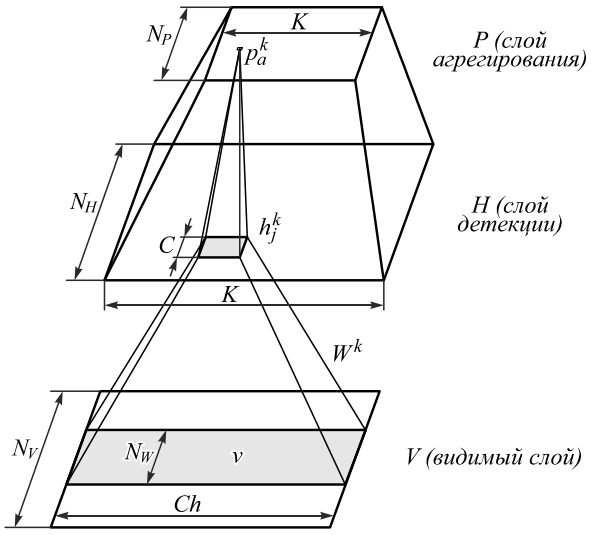

Рис. 1. Структура свёрточной ограниченной машины Больцмана

Представим, что входной слой сети состоит из матрицы вещественных нейронов размерности $N_V \times Ch$, где N_V – количество окон, на которые разбивается аудиосигнал, Ch – количество каналов спектра. Для создания скрытого слоя используются K фильтров размерности $N_W \times Ch$ с весами W^K, которые также называют базами. Слой детекции H состоит из матрицы размерности $N_H \times K$ с нейронами, разделяющими общие веса W^k ($N_H = N_V - N_W + 1$). Также для каждой группы фильтров задаётся общее смещение b_k и общее смещение для видимого слоя c.

Слой агрегирования P состоит из матрицы вещественных нейронов размерности $N_P \times K$. Для $k \in \{1, ..., K\}$ слой агрегирования P уменьшает размерность признаков слоя детекции H в C раз, таким образом, $N_P = N_H / C$. То есть слой детекции H разбивается на сегменты размерности $C \times 1$ и каждый такой блок a соединён с одним нейроном p_a^k в слое агрегирования. В качестве значения нейрона p_a^k случайным образом выбирается либо значение одного из нейронов в слое детекции H, либо значение нейрона обнуляется.

Тогда функцию энергии СОМБ (5) можно задать как [4]:

$$E(v,h) = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k W_r^k v_{j+r-1} - \\ - \sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i . \quad (5)$$

Зададим совместные и условные функции распределения вероятностей для данной сети (2)–(4):

$$P(v,h) = \frac{1}{z}\exp(-E(v,h)), \quad (6)$$

$$P(h_j^k = 1 \mid v) = sigmoid((\tilde{W}_j^{\,k} *_v v)_j + b_k), \quad (7)$$

$$P(v_i \mid h) = Normal(\sum_k (\tilde{W}^k *_f h^k)_i + c, 1), \quad (8)$$

где $*_v$ – действительная свёртка, $*_f$ – полная свёртка,

$\tilde{W}_j^{\,k} \triangleq W_{n_W - j + 1}^k$ [4].

Для заданного вектора размерности m и ядра размерности n, где $m > n$, валидная свёртка даёт $(m - n + 1)$-мерный вектор, полная свертка даёт $(m + n - 1)$-мерный вектор. Так как все нейроны скрытого слоя условно независимы от другого слоя, вывод в сети можно осуществить, используя семплирование по Гиббсу.

Свёрточная глубокая сеть доверия представляет собой композицию простых свёрточных ограниченных машин Больцмана, благодаря чему скрытый слой каждой подсети служит видимым слоем для следующей. Из-за этого можно произвести быструю послойную процедуру обучения без учителя, в которой для оценки градиента относительное расхождение [25] применяет-

ся к каждой подсети по очереди, начиная с первой пары слоёв. На видимый слой сети подаются данные из обучающего набора, последующие скрытые слои принимают на вход данные с выхода предыдущих.

Примем слой детекции H и слой агрегирования P в качестве скрытого слоя H' и зададим общий алгоритм обучения СГСД следующим образом:

1. Представить два нижних слоя (входной и первый скрытый) как СОМБ. Произвести обучение СОМБ входных данных из видимого слоя **V** и получить матрицу её весовых коэффициентов **W**, которая будет описывать связи между двумя нижними слоями сети.
2. Произвести вычисления, пропустив через уже обученную СОМБ входные данные **V**, и получить данные скрытого слоя **H'** на выходе после активации узлов первого скрытого слоя.
3. Повторять шаги 1 и 2 для всех последующих пар слоев СГСД, используя в качестве входных данных выходы предыдущего слоя **H'** до тех пор, пока не будут обучены все слои СГСД.

4. Экспериментальная часть исследования

Для оценки точности систем верификации диктора с применением СГСД были использованы две различных метрики: равная ошибка 1-го и 2-го рода (*Equal Error Rate*, EER) и минимальная функция стоимости обнаружения (*minimum detection cost function*, minDCF) с параметрами NIST SRE 2008. Равная ошибка первого и второго рода определяет ошибку распознавания диктора при условии равенства вероятности пропуска самозванца и отказа законному пользователю. Данная характеристика используется как для оценки текстозависимых, так и текстонезависимых систем верификации диктора.

Под точностью верификации будем понимать отношение количества верно распознанных целевых дикторов и самозванцев по отношению к общему количеству проведённых тестов. Точность верификации диктора Acc с равной ошибкой 1-го и 2-го рода EER можно связать следующим образом:

$$Acc = 1 - (2 * \mathrm{EER}). \quad (9)$$

Для проведения экспериментов было необходимо задать структуру и параметры СГСД, а после этого произвести обучение данной сети. Была использована следующая структура свёрточной глубокой сети доверия: сеть состояла из 3 слоёв, первый и второй слой состоял из 300 баз, третий – из 60. Входной слой состоял из 80 нейронов ($Ch = 80$). На вход сети были поданы спектрограммы аудиозаписей, полученные с уменьшением размерности с помощью метода главных компонент. Данные, подаваемые на видимый слой, выбирались окнами по 20 мс со смещением 10 мс. Для каждой базы в скрытых слоях была использована размерность фильтра $N_W = 6$, коэффициент агрегирования $C = 3$. Параметры для первого и второго слоя сети были взяты из [4]. Параметры для третьего слоя выбраны авторами самостоятельно.

В результате обучения СГСД были получены три обученных слоя сети, выходы каждого из которых можно использовать в качестве признаков для проведения верификации диктора. Для оценки системы верификации с полученными признаками выходы СГСД были поданы на вход модели ГС, состоящей из 256 компонентов. Модели дикторов были получены с помощью MAP-адаптации, с адаптацией только векторов математических ожиданий и фактором релевантности $r = 10$. GMM-UBM-система, которая рассматривалась в роли базовой системы верификации, была создана с применением библиотеки MSR Identity Toolbox [26].

Помимо ГС, в качестве классификаторов, использующих признаки, полученные из СГСД, были использованы машина опорных векторов (SVM), алгоритм ансамблевой классификации AdaBoost M1 и классификатор, основанный на линейном дискриминантном анализе (*Linear Discriminant Analysis*, LDA). Для обучения данных классификаторов использовался подход «один против всех». При этом данные, использованные для обучения УФМ, не были включены в обучающую выборку.

Помимо оценки точности отдельных классификаторов, был создан ансамбль классификаторов, использующий как классификаторы с традиционными речевыми признаками, так и классификаторы с признаками, извлечёнными из СГСД. При создании ансамбля классификаторов использовались классификаторы для всех трёх типов признаков СГСД. Для итоговой оценки тестовой аудиозаписи была использована взвешенная сумма выходов нескольких классификаторов. Для определения весов был использован генетический алгоритм.

Экспериментальное исследование было проведено на двух речевых корпусах: собственном речевом корпусе, включающем записи речи 25 дикторов-мужчин и 25 женщин, и речевом корпусе TIMIT, включающем в себя записи речи 630 дикторов.

Собственный речевой корпус содержит записи произнесённых без предварительной подготовки предложений на русском языке, взятых из художественной литературы, или поговорок. Суммарная длина записей речи для каждого диктора составляет не менее 6 мин, включая 50 сегментов различной длины. Каждый диктор был записан на микрофон в условиях нешумной аудитории. Аудиозаписи имеют следующие параметры: частота дискретизации – 8000 Гц, разрядность – 16 бит.

Весь речевой корпус, состоящий из записей речи 50 дикторов, был разделён на две части – одна для обучения УФМ (состоит из записей 30 дикторов), вторая – для обучения и тестирования моделей дикторов (состоит из записей оставшихся 20 дикторов).

Обе части включают в себя равное количество дикторов мужчин и женщин.

Для MAP-адаптации моделей дикторов использовались 40 речевых сигналов. Оставшиеся 10 сигналов каждого диктора применялись для тестирования системы верификации. В сумме было произведено 4 000 тестов для каждого набора признаков, по 10 положительных (тестируется целевой диктор) и 190 отрицательных (тестируется диктор-нарушитель) для каждого диктора.

Рассмотрим некоторые из полученных результатов оценок точности полученных классификаторов с применением признаков, извлечённых из различных слоёв СГСД, для данного речевого корпуса (табл. 1). Каждый из использованных классификаторов проигрывает по точности по сравнению с применением 14 мел-кепстральных коэффициентов в качестве речевых признаков. Для большинства классификаторов ошибка EER была получена примерно в 2 раза большая, чем при использовании традиционных речевых признаков.

Табл. 1. Сравнение точности классификаторов, использующих признаки, извлечённые из различных слоёв СГСД

Тип классификатора	% EER	minDCF*100
ГС, СГСД1	2,00	1,00
ГС, СГСД2	3,50	1,74
ГС, СГСД3	10,00	5,77
LDA, СГСД1	2,00	1,42
LDA, СГСД2	1,97	1,46
LDA, СГСД3	7,08	4,86
ГС, 14 МКК	**1,00**	**0,93**

Однако совмещение в ансамбле классификаторов, использующих традиционные речевые признаки и признаки, извлечённые из СГСД, дало уменьшение EER. Итоговый ансамбль, показавший наименьшую ошибку EER = 0,21 % (рис. 2), состоит из следующих классификаторов: Гауссовой смеси, обученной на векторе признаков, который был получен с помощью жадного алгоритма добавления-удаления; классификаторов LDA и AdaBoost M1, обученных на признаках с первого скрытого слоя СГСД; машины опорных векторов SVM, обученной на признаках из третьего скрытого слоя СГСД. С учётом доверительного уровня 95 % доверительный интервал для оценки равной ошибки 1-го и 2-го рода составляет EER = 0,21 ± 0,14 %.

Вторая часть исследований была проведена на речевом корпусе TIMIT [27], который часто используется в работах, связанных с обработкой и распознаванием речи [28 – 31]. В данном корпусе содержатся аудиозаписи носителей восьми основных диалектов американского английского языка, для каждого из которых имеется по десять фонетически разнообразных, отражающих диалект или фонетически компактных фраз, часть из которых одинаковы для всех дикторов, часть отличается друг от друга. Всего в корпусе имеются аудиозаписи 438 дикторов-мужчин и 192 дикторов-женщин.

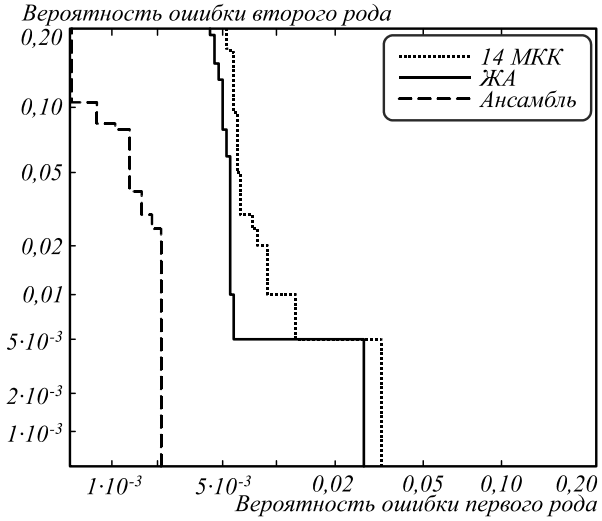

Рис. 2. Кривые компромиссного определения ошибки (DET кривые), полученные при исследовании собственного речевого корпуса

Наибольший вклад в финальную оценку тестовой аудиозаписи дали в порядке убывания: классификатор на ГС, классификатор LDA, классификатор SVM, классификатор AdaBoost M1. Следует отметить, что в повышении точности также участвовал третий слой СГСД, который выделяет признаки более высокого уровня. После того, как данный классификатор был убран из ансамбля, произошло увеличение ошибки EER до 0,5 %. Отсюда следует, что СГСД позволяет выделять признаки, повышающие точность системы верификации диктора по произвольной фразе (табл. 2).

Табл. 2. Сравнение точности ансамблей классификаторов и Гауссовой смеси с применением традиционных и СГСД-признаков

Тип классификатора	% EER	minDCF*100
Ансамбль	**0,21**	**0,21**
Ансамбль, без SVM3	0,50	0,52
ГС, ЖА	0,58	0,62
ГС, 14 МКК	1,00	0,93

Для сравнения полученных результатов с работой [31] были использованы следующие условия проведения эксперимента: для обучения фоновой модели были использованы аудиозаписи 530 дикторов, оставшиеся аудиозаписи 100 дикторов использованы на стадии тестирования, где 30 дикторов – женщины и 70 дикторов – мужчины. Для обучения фоновой модели были использованы все 5300 аудиозаписей, для обучения моделей отдельных дикторов были использованы 9 из 10 аудиозаписей, оставшаяся одна запись была использована для тестирования. Верификационные испытания состояли из 10000 тестов, из которых 100 тестов положительных, в этом случае тестируемая аудиозапись соответствует модели дик-

тора, и 9900 отрицательных тестов, когда на аудиозаписи отсутствует голос целевого диктора.

Для оценки точности полученной системы верификации был взят ансамбль классификаторов со структурой, аналогичной использованной ранее (табл. 3). Как и в случае с собственным речевым корпусом, наибольшую точность показал ансамбль классификаторов, элементы которого используют как традиционные речевые признаки, так и признаки, извлечённые с помощью СГСД (рис. 3). Также можно отметить влияние высокоуровневых признаков, извлечённых из третьего слоя СГСД, на формирование итогового результата. Отсутствие данного классификатора в ансамбле увеличивает равную ошибку 1-го и 2-го рода до 0,32 %. С учётом доверительного уровня 95 %, доверительный интервал для оценки равной ошибки 1-го и 2-го рода составляет EER = 0,23 ± 0,09 %.

Табл. 3. Сравнение точности систем верификации диктора различных типов для речевого корпуса TIMIT

Тип классификатора	% EER	minDCF*100
Ансамбль	**0,23**	**0,23**
Ансамбль, без SVM3	0,32	0,32
ГС, ЖА	0,56	0,55
ГС, 13 МКК	1,47	1,37
i-vector, ЖА	3,00	1,18
i-vector, 13 МКК	1,14	1,01
Meriem et al. (ГС, 13 МКК)	0,58	0,52

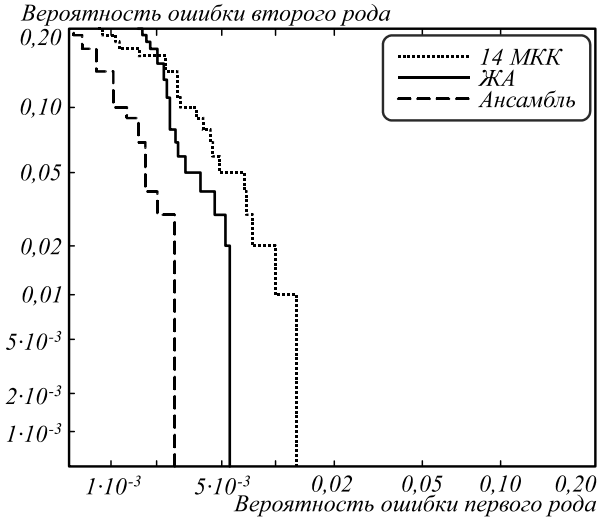

Рис. 3. Кривые компромиссного определения ошибки (DET кривые), полученные при исследовании речевого корпуса TIMIT

Проводя сравнение результатов с работой [31], можно отметить уменьшение ошибки EER с 0,58 % до 0,23 % с применением предложенного ансамбля классификаторов. Система верификации, основанная на построении *i*-векторов, показала меньшую точность по сравнению с другими типами систем. Также применение вектора признаков, полученного с помощью жадного алгоритма отбора признаков, увеличило EER в системе, основанной на применении *i*-векторов, в отличие от GMM-UBM-системы. Несмотря на попытку полностью повторить условия проведения эксперимента в работе [31], авторами были получены значительно отличающиеся результаты для системы верификации, основанной на использовании ГС и 13 мел-кепстральных коэффициентов.

5. Анализ обученной СГСД

В результате обучения СГСД были получены три обученных слоя сети, выходы каждого из которых можно использовать в качестве признаков для проведения верификации диктора. Рассмотрим полученные выходы сети на примере фразы «Со спокойным мужеством Скайлс ожидал всего в этом безумном городе», произнесённой мужчиной и женщиной (рис. 4 – 7).

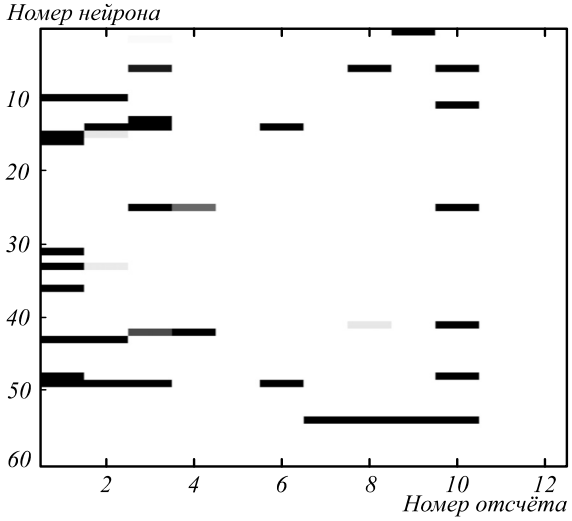

Рис. 4. Выходные значения нейронов третьего скрытого слоя обученной СГСД для фразы диктора - мужчины

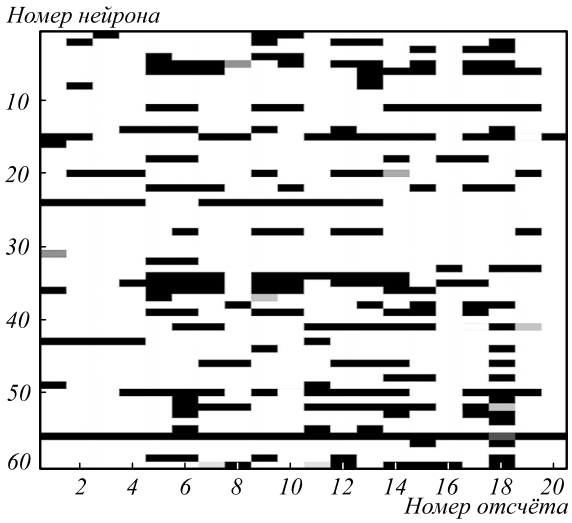

Рис. 5. Выходные значения нейронов третьего скрытого слоя обученной СГСД для фразы диктора-женщины

Исходя из визуального представления функций активации нейронов третьего скрытого слоя СГСД, видны сильные отличия – для диктора-женщины ак-

тивируется гораздо больше нейронов (рис. 4, 5). Аналогично можно отметить большее количество активаций фильтров второго скрытого слоя СГСД для фразы диктора-женщины, чем для фразы диктора-мужчины (рис. 6, 7). Данные отличия наблюдались не только для двух данных дикторов, но и для других пар дикторов.

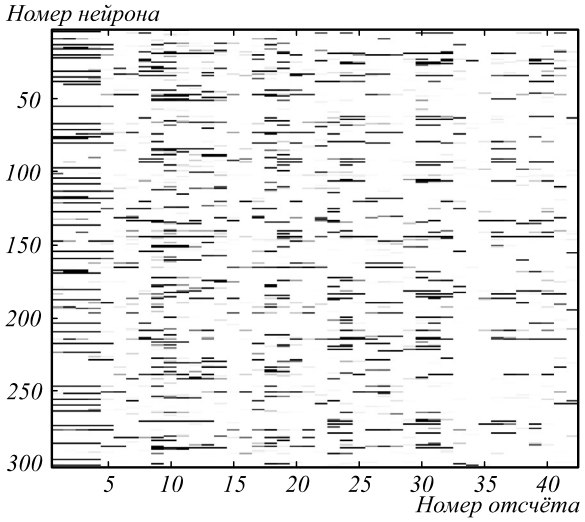

Рис. 6. Выходные значения нейронов второго скрытого слоя обученной СГСД для фразы диктора-мужчины

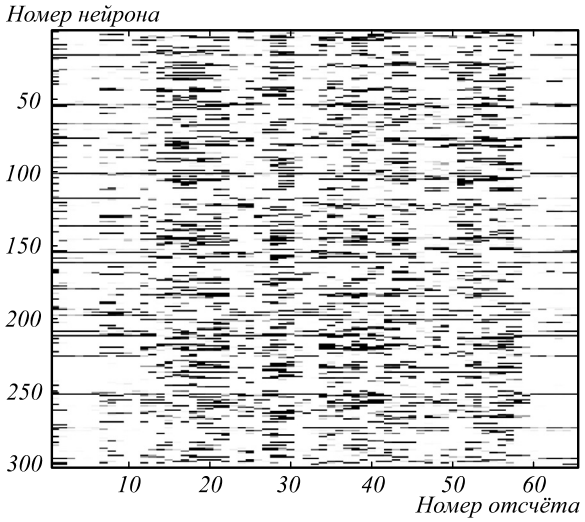

Рис. 7. Выходные значения нейронов второго скрытого слоя обученной СГСД для фразы диктора-женщины

Следовательно, можно сделать вывод, что фильтры в скрытых слоях СГСД обучаются разделять речь дикторов-мужчин и дикторов-женщин. Одно из возможных применений данной особенности признаков, извлеченных из СГСД, – решение задачи идентификации пола диктора.

На рис. 8 изображены паттерны трёх случайным образом взятых фильтров первого скрытого слоя обученной СГСД. Можно увидеть, что после обучения данные фильтры пытаются найти определённые спектральные шаблоны, соответствующие различным голосам или фонемам. Видно, что более тёмные области в спектре фильтра соответствуют местам концентрации энергии различных типов звуков.

На более высоких уровнях на основе фильтров первого уровня строится высокоуровневое представление речи дикторов. Эти высокоуровневые представления являются спектральными шаблонами, объединяющими в себе частотные и временные характеристики акустических сигналов. Можно сделать предположение, что вполне возможным является применение данных признаков для распознавания речи или идентификации пола говорящего.

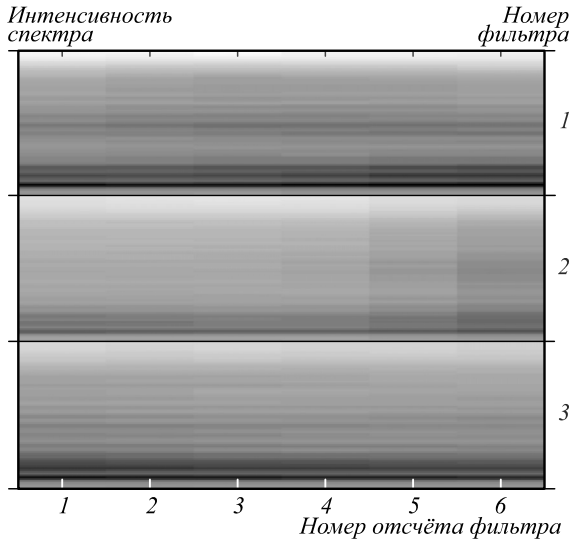

Рис. 8. Паттерны трёх случайным образом взятых фильтров первого скрытого слоя обученной СГСД

Заключение

В данной работе было рассмотрено применение свёрточной глубокой сети доверия для извлечения из аудиозаписей речевых признаков, используемых для решения задачи верификации диктора по произвольной фразе. Метод извлечения признаков отличается от аналогичных решений расширенной архитектурой нейронной сети, выделяющей признаки высокого уровня и уменьшающей их общее количество.

Точность предложенных методов верификации была проверена на двух речевых корпусах: собственном речевом корпусе, включающем аудиозаписи 50 дикторов, и речевом корпусе TIMIT, включающем аудиозаписи 630 дикторов. Непосредственное применение признаков, извлечённых с помощью СГСД, не дало увеличения точности по сравнению с использованием традиционных речевых признаков. Однако применение данных признаков в составе ансамбля классификаторов позволило достичь уменьшения равной ошибки 1-го и 2-го рода до 0,21 % на собственном речевом корпусе и до 0,23 % на речевом корпусе TIMIT.

Анализ обученной СГСД показал, что данная сеть позволяет находить и выделять спектральные шаблоны в спектре речи. Высокоуровневые признаки зна-

чительно отличаются для голосов дикторов разного пола. Сделано предположение, что возможно применение данных высокоуровневых признаков для решения других задач обработки речи – определения пола говорящего и распознавания речи.

Благодарности

Результаты были получены в рамках выполнения базовой части государственного задания Минобрнауки России, проект 8.9628.2017/8.9.

Литература

1. **Campbell, J.P.** Speaker recognition: a tutorial / J.P. Campbell // Proceedings of the IEEE. – 1997. – Vol. 85, Issue 9. – P. 1437-1462.
2. **Soldatova, O.P.** Convolutional neural network applied to handwritten digits recognition / O.P. Soldatova, A.A. Garshin // Computer Optics. – 2010. – Vol. 34, Issue 2. – P. 252-259.
3. **Lee, H.** Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations / H. Lee, R. Grosse, R. Ranganath, A.Y. Ng // Proceedings of the 26th Annual International Conference on Machine Learning. – 2009. – P. 609-616.
4. **Lee, H.** Unsupervised feature learning for audio classification using convolutional deep belief networks / H. Lee, P. Pham, Y. Largman, A.Y. Ng // Advances in Neural Information Processing Systems. – 2009. – P. 1096-1104.
5. **Ren, Y.** Convolutional deep belief networks for feature extraction of EEG signal / Y. Ren, Y. Wu // 2014 International Joint Conference on Neural Networks (IJCNN). – 2014. – P. 2850-2853.
6. **Sahidullah, M.** A novel windowing technique for efficient computation of MFCC for speaker recognition / M. Sahidullah, G. Saha // IEEE Signal Processing Letters. – 2013. – Vol. 20, Issue 2. – P. 149-152.
7. **Motlicek, P.** Employment of subspace gaussian mixture models in speaker recognition / P. Motlicek, S. Dey, S. Madikeri, L. Burget // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2015. – P. 4445-4449.
8. **Greenberg, C.S.** The NIST 2014 speaker recognition i-vector machine learning challenge / C.S. Greenberg, D. Bansé, G.R. Doddington, D. Garcia-Romero, J.J. Godfrey, T. Kinnunen, A.F. Martin, A. McCree, M. Przybocki, D.A. Reynolds // Odyssey: The Speaker and Language Recognition Workshop. – 2014. – P. 224-230.
9. **Lei, Y.** A novel scheme for speaker recognition using a phonetically-aware deep neural network / Y. Lei, N. Scheffer, L. Ferrer, M. McLaren // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2014. – P. 1695-1699.
10. **Stafylakis, T.** Compensation for phonetic nuisance variability in speaker recognition using DNNs / T. Stafylakis, P. Kenny, V. Gupta, J. Alam, M. Kockmann // Odyssey: The Speaker and Language Recognition Workshop. – 2016. – P. 340-345.
11. **Kenny, P.** Deep neural networks for extracting baum-welch statistics for speaker recognition / P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, J. Alam // Proceedings of the Odyssey. – 2014. – P. 293-298.
12. **Xu, L.** Rapid Computation of I-vector / L. Xu, K.A. Lee, H. Li, Z. Yang // Odyssey: The Speaker and Language Recognition Workshop. – 2016. – P. 47-52.
13. **McLaren, M.** Exploring the role of phonetic bottleneck features for speaker and language recognition / M. McLaren, L. Ferrer, A. Lawson // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2016. – P. 5575-5579.
14. **Richardson, F.** Deep neural network approaches to speaker and language recognition / F. Richardson, D. Reynolds, N. Dehak // IEEE Signal Processing Letters. – 2015. – Vol. 22, Issue 10. – P. 1671-1675.
15. **Reynolds, D.A.** Speaker verification using adapted Gaussian mixture models / D.A. Reynolds, T.F. Quatieri, R.B. Dunn // Digital Signal Processing. – 2000. – Vol. 10, Issue 1. – P. 19-41.
16. **Sizov, A.** Joint speaker verification and antispoofing in the I-vector space / A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, S. Marcel // IEEE Transactions on Information Forensics and Security. – 2015. – Vol. 10, Issue 4. – P. 821-832.
17. **Variani, E.** Deep neural networks for small footprint text-dependent speaker verification / E. Variani, X. Lei, E. McDermott, I.L. Moreno, J. Gonzalez-Dominguez // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2014. – P. 4052-4056.
18. **Jung, J.W.** A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result / J.W. Jung, H.S. Heo, I.H. Yang, H.J. Shim, H.J. Yu // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – P. 5349-5353.
19. **Rohdin, J.** End-to-end DNN based speaker recognition inspired by i-vector and PLDA / J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, L. Burget // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2018. – P. 4874-4878.
20. **Рахманенко, И.А.** Анализ идентификационных признаков в речевых данных с помощью GMM-UBM системы верификации диктора / И.А. Рахманенко, Р.В. Мещеряков // Труды СПИИРАН. – 2017. – Т. 52, № 3. – С. 22-50.
21. **Davis, S.B.** Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences / S.B. Davis, P. Mermelstein // IEEE Transactions on Acoustics, Speech and Signal Processing. – 1980. – Vol. 28, Issue 4. – P. 357-366.
22. **Jurafsky, D.** Speech and language processing / D. Jurafsky, J.H. Martin. – 2nd ed. – New Jersey: Pearson Education, 2009. – 1026 p.
23. **Eyben, F.** Recent developments in opensmile, the munich open-source multimedia feature extractor / F. Eyben, F. Weninger, F. Gross, B. Schuller // Proceedings of the 21st ACM International Conference on Multimedia. – 2013. – P. 835-838.
24. **Hinton, G.E.** A fast learning algorithm for deep belief nets / G.E. Hinton, S. Osindero, Y.W. Teh // Neural Computation. – 2006. – Vol. 18, Issue 7. – P. 1527-1554.
25. **Hinton, G.E.** Training products of experts by minimizing contrastive divergence / G.E. Hinton // Neural Computation. – 2002. – Vol. 14, Issue 8. – P. 1771-1800.
26. **Sadjadi, S.O.** MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research / S.O. Sadjadi, M. Slaney, L. Heck // Speech and Language Processing Technical Committee Newsletter. – 2013. – Vol. 1, Issue 4. – P. 1-32.
27. **Zue, V.** Speech database development at MIT: TIMIT and beyond / V. Zue, S. Seneff, J. Glass // Speech Communication. – 1990. – Vol. 9, Issue 4. – P. 351-356.
28. **Yoshimura, T.** Discriminative feature extraction based on sequential variational autoencoder for speaker recognition /

T. Yoshimura, N. Koike, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda // 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). – 2018. – P. 1742-1746.
29. **Zeng, C.Y.** Stacked autoencoder networks based speaker recognition / C.Y. Zeng, C.F. Ma, Z.F. Wang, J.X. Ye // 2018 International Conference on Machine Learning and Cybernetics (ICMLC). – 2018. – Vol. 1. – P. 294-299.
30. **Chorowski, J.K.** Attention-based models for speech recognition / J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio // Advances in Neural Information Processing Systems. – 2015. – P. 577-585.
31. **Meriem, F.** Robust speaker verification using a new front end based on multitaper and gammatone filters / F. Meriem, H. Farid, B. Messaoud, A. Abderrahmene // 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems. – 2014. – P. 99-103.

Сведения об авторах

Рахманенко Иван Андреевич, 1990 года рождения, кандидат технических наук, в 2012 году окончил Томский государственный университет систем управления и радиоэлектроники (ТУСУР) по специальности «Комплексное обеспечение информационной безопасности автоматизированных систем», в 2017 году защитил кандидатскую диссертацию в Томском государственном университете систем управления и радиоэлектроники. Работает доцентом на кафедре безопасности информационных систем, ТУСУР. Область научных интересов: верификация диктора, обработка речи, машинное обучение, программно-аппаратные средства защиты информации. E-mail: *ria@keva.tusur.ru* .

Шелупанов Александр Александрович, 1954 года рождения, доктор технических наук, профессор. Работает заведующим кафедрой комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС), ТУСУР. Область научных интересов: информационная безопасность, программно-аппаратные средства защиты информации, построение систем аутентификации. E-mail: *saa@keva.tusur.ru* .

Костюченко Евгений Юрьевич, 1983 года рождения, кандидат технических наук, доцент, в 2005 году окончил Томский государственный университет систем управления и радиоэлектроники (ТУСУР) по специальности «Комплексное обеспечение информационной безопасности автоматизированных систем», в 2010 году защитил кандидатскую диссертацию в Томском политехническом университете. Работает ведущим научным сотрудником лаборатории медико-биологических исследований (ЛМБИ) и доцентом на кафедре КИБЭВС, ТУСУР. Область научных интересов: обработка речи, машинное обучение, системы аутентификации, оценка качества речи. E-mail: *key@keva.tusur.ru* .

ГРНТИ: 28.23.15
Поступила в редакцию 20 августа 2019 г. Окончательный вариант – 13 октября 2019 г.

Automatic text-independent speaker verification using convolutional deep belief network

I.A. Rakhmanenko [1], A.A. Shelupanov [1], E.Y. Kostyuchenko [1]
[1] Tomsk State University of Control Systems and Radioelectronics, prospect Lenina 40, 634050, Tomsk, Russia

Abstract

This paper is devoted to the use of the convolutional deep belief network as a speech feature extractor for automatic text-independent speaker verification. The paper describes the scope and problems of automatic speaker verification systems. Types of modern speaker verification systems and types of speech features used in speaker verification systems are considered. The structure and learning algorithm of convolutional deep belief networks is described. The use of speech features extracted from three layers of a trained convolution deep belief network is proposed. Experimental studies of the proposed features were performed on two speech corpora: own speech corpus including audio recordings of 50 speakers and TIMIT speech corpus including audio recordings of 630 speakers. The accuracy of the proposed features was assessed using different types of classifiers. Direct use of these features did not increase the accuracy compared to the use of traditional spectral speech features, such as mel-frequency cepstral coefficients. However, the use of these features in the classifiers ensemble made it possible to achieve a reduction of the equal error rate to 0.21% on 50-speaker speech corpus and to 0.23% on the TIMIT speech corpus.

Keywords: speaker recognition, speaker verification, Gaussian mixture models, GMM-UBM system, speech features, speech processing, deep learning, neural networks, pattern recognition.

Citation: Rakhmanenko IA, Shelupanov AA, Kostyuchenko EYu. Automatic text-independent speaker verification using convolutional deep belief network. Computer Optics 2020; 44(4): 596-605. DOI: 10.18287/2412-6179-CO-621.

Acknowledgements: The work was funded within the basic part of the government project of the Russian Federation Education and Science Ministry, project 8.9628.2017/8.9.

References

[1] Campbell JP. Speaker recognition: a tutorial. Proc IEEE Inst Electr Electron Eng 1997; 85(9): 1437-1462.

[2] Soldatova OP, Garshin AA. Convolutional neural network applied to handwritten digits recognition. Computer Optics 2010; 34(2): 252-259.

[3] Lee H, Grosse R, Ranganath R, Ng AY. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proc 26th Annual International Conference on Machine Learning 2009: 609-616.

[4] Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. Adv Neural Inform Process Syst 2009: 1096-1104.

[5] Ren Y, Wu Y. Convolutional deep belief networks for feature extraction of EEG signal. IJCNN 2014: 2850-2853.

[6] Sahidullah M, Saha G. A novel windowing technique for efficient computation of MFCC for speaker recognition. IEEE Signal Process Lett 2013; 20(2): 149-152.

[7] Motlicek P, Dey S, Madikeri S, Burget L. Employment of subspace gaussian mixture models in speaker recognition. ICASSP 2015: 4445-4449.

[8] Greenberg CS, Bansé D, Doddington GR, Garcia-Romero D, Godfrey JJ, Kinnunen T, Martin AF, McCree A, Przybocki M, Reynolds DA. The NIST 2014 speaker recognition i-vector machine learning challenge. Odyssey: The Speaker and Language Recognition Workshop 2014: 224-230.

[9] Lei Y, Scheffer N, Ferrer L, McLaren M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. ICASSP 2014: 1695-1699.

[10] Stafylakis T, Kenny P, Gupta V, Alam J, Kockmann M. Compensation for phonetic nuisance variability in speaker recognition using DNNs. Odyssey: The Speaker and Language Recognition Workshop 2016: 340-345.

[11] Kenny P, Gupta V, Stafylakis T, Ouellet P, Alam J. Deep neural networks for extracting baum-welch statistics for speaker recognition. Proc Odyssey 2014: 293-298.

[12] Xu L, Lee KA, Li H, Yang Z. Rapid Computation of I-vector. Odyssey: The Speaker and Language Recognition Workshop 2016: 47-52.

[13] McLaren M, Ferrer L, Lawson A. Exploring the role of phonetic bottleneck features for speaker and language recognition. ICASSP 2016: 5575-5579.

[14] Richardson F, Reynolds D, Dehak N. Deep neural network approaches to speaker and language recognition. IEEE Signal Process Lett 2015; 22(10): 1671-1675.

[15] Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digit Signal Process 2000; 10(1): 19-41.

[16] Sizov A, Khoury E, Kinnunen T, Wu Z, Marcel S. Joint speaker verification and antispoofing in the I-vector space. IEEE Trans Inf Forensics Secur 2015; 10(4): 821-832.

[17] Variani E, Lei X, McDermott E, Moreno IL, Gonzalez-Dominguez J. Deep neural networks for small footprint text-dependent speaker verification. ICASSP 2014: 4052-4056.

[18] Jung JW, Heo HS, Yang IH, Shim HJ, Yu HJ. A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. ICASSP 2018: 5349-5353.

[19] Rohdin J, Silnova A, Diez M, Plchot O, Matějka P, Burget L. End-to-end DNN based speaker recognition inspired by i-vector and PLDA. ICASSP 2018: 4874-4878.

[20] Rakhmanenko IA, Meshcheryakov RV. Identification features analysis in speech data using GMM-UBM speaker verification system [In Russian]. SPIIRAS Proc 2017; 52(3): 22-50.
[21] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Audio Speech Lang Process 1980; 28(4): 357-366.
[22] Jurafsky D, Martin JH. Speech and language processing. 2nd ed. New Jersey: Pearson Education; 2009.
[23] Eyben F, Weninger F, Gross F, Schuller B. Recent developments in opensmile, the munich open-source multimedia feature extractor. Proc 21st ACM Int Conf Multimedia 2013: 835-838.
[24] Hinton GE, Osindero S, The YW. A fast learning algorithm for deep belief nets. Neural Comput 2006; 18(7): 1527-1554.
[25] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Comput 2002; 14(8): 1771-1800.
[26] Sadjadi SO, Slaney M, Heck L. MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research. Speech and Language Processing Technical Committee Newsletter 2013; 1(4): 1-32.
[27] Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond. Speech Commun 1990; 9(4): 351-356.
[28] Yoshimura T, Koike N, Hashimoto K, Oura K, Nankaku Y, Tokuda K. Discriminative feature extraction based on sequential variational autoencoder for speaker recognition. APSIPA ASC 2018: 1742-1746.
[29] Zeng CY, Ma CF, Wang ZF, Ye JX. Stacked Autoencoder Networks Based Speaker Recognition. ICMLC 2018; 1: 294-299.
[30] Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech. Advances in neural information processing systems 2015: 577-585.
[31] Meriem F, Farid H, Messaoud B, Abderrahmene A. Robust speaker verification using a new front end based on multitaper and gammatone filters. 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems 2014: 99-103.

Authors' information

Ivan Andreevich Rakhmanenko (b. 1990) graduated from Tomsk State University of Control Systems and Radioelectronics (TUSUR) in 2012, majoring in Complex Information Security of Automated Systems. Defended PhD thesis in 2017. Currently he works as the assistant professor at the Information Systems Security chair of TUSUR. Research interests are speaker verification, speech processing, machine learning, hardware and software information protection systems. E-mail: *ria@keva.tusur.ru* .

Alexander Alexandrovich Shelupanov (b. 1954) Doctor of Technical Sciences, Professor. Currently he works as the head of Complex Information Security of Electronic Computing Systems department (KIBEVS), TUSUR. Research interests: information security, software and hardware information protection, building authentication systems. E-mail: *saa@keva.tusur.ru* .

Evgeny Yurievich Kostyuchenko (b. 1983) Candidate of Technical Sciences, Associate Professor. In 2005 graduated from Tomsk State University of Control Systems and Radio Electronics (TUSUR) majoring in Complex Information Security of Automated Systems, in 2010 he defended his PhD thesis at Tomsk Polytechnic University. He works as a leading researcher at the Laboratory for Biomedical Research (LMBI) and associate professor of KIBEVS department, TUSUR. Research interests: speech processing, machine learning, authentication systems, speech quality assessment. E-mail: *key@keva.tusur.ru* .

Received August 20, 2019. The final version – October 13, 2019.