ПРИМЕНЕНИЕ ПРЕОБРАЗОВАНИЯ РАССЕЯНИЯ НА КОЭФФИЦИЕНТАХ ДИСКРЕТНОГО ВЕЙВЛЕТ-РАЗЛОЖЕНИЯ К ЗАДАЧЕ БИОМЕТРИЧЕСКОЙ ВЕРИФИКАЦИИ ДИКТОРОВ

Лепендин А.А., Гапонов Д.А., Филин Я.А., Ладыгин П.С.
Алтайский государственный университет, г. Барнаул
email: andrey.lependin@gmail.com

Аннотация. В работе предложен новый подход к вычислению информативных признаков речевого сигнала для задачи верификации диктора. К сигналу применялось многоуровневое преобразование, вычисляющее коэффициенты рассеяния на основе дискретного вейвлет-разложения. Полученные вектора признаков использовались в качестве входных данных нейронной сети с временными задержками. На их основе нейронной сетью вычислялись вектора идентичности дикторов, которые непосредственно применялись для биометрической верификации. Предложенный подход был апробирован на данных из наборов голосовых образцов VoxCeleb1 и VoxCeleb2. Была показана его эффективность в сравнении с существующими методами верификации на основе глубоких нейронных сетей.

Ключевые слова: голосовая верификация, дискретное вейвлет-преобразование, преобразование рассеяния, нейронная сеть с временными задержками, вектор идентичности диктора.

Одной из критически важных задач в приложениях, где необходим повышенный уровень безопасности, являются обработка и управление доступом к ресурсам. В настоящее время наблюдается растущая тенденция по использованию биометрических данных вместо смарт-карт или ключей, для контроля доступа и идентификации [1]. Среди всех биометрических характеристик наиболее удобной является голос, так как пользователи в основном чувствуют себя комфортно из-за простоты использования. Речевая биометрия с другой стороны относительно больше подвержена возможным атакам на биометрическое предъявление [2]. Для эффективного отслеживания подобных атак, необходима разработка новых подходов к биометрической верификации дикторов.

В данной работе для верификации пользователей предложено использовать в качестве информативных признаков речевых аудиосигналов разложения рассеяния (scattering decomposition), которые представляет собой иерархические спектральные разложения на основе банков вейвлет-подобных фильтров [3]. Данные признаки использовались для обучения нейронной сети, извлекающей вектора идентичности дикторов, используемые непосредственно для биометрической верификации.

Основными преимуществами преобразования рассеяния перед «классическими» спектральными разложениями является его инвариантность к слабым локальным искажениям сигнала как в частотной, так и во временной области, а также сохранение полной информации о сигнале, что, в принципе, позволяет проводить полное с точностью до фазы восстановление исходного сигнала [4].

Рассмотрим подробнее схему извлечения коэффициентов разложения рассеяния. Они группируются по так называемым уровням разложения сигнала. Каждый новый уровень вычисляется путем пропускания результатов разложения предыдущего уровня с помощью банка дискретных вейвлет-фильтров и вычисления абсолютной величины усредненных результатов фильтрации (рисунок 1). В качестве начального (нулевого) уровня выступал исходный аудиосигнал.

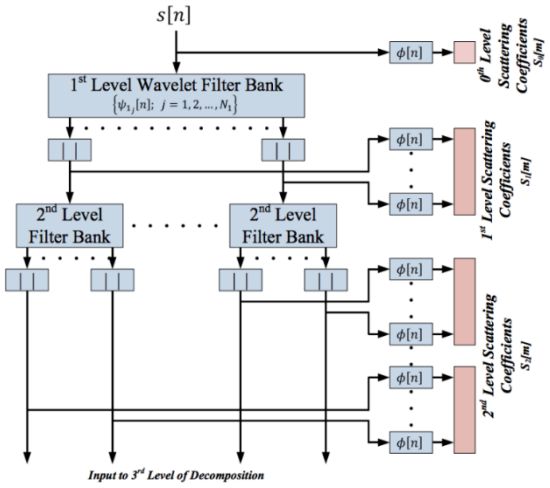

Рисунок 1. Схема разложения рассеяния (из [3]).

На k-м уровне разложения банк из N_k фильтров представлял собой вейвлет-преобразование, заданное набором вейвлетов $\psi_{k_j}[n]$ вида:

$$\psi_{kj}[n] = \frac{1}{a}\psi_k\left[\frac{nT}{a}\right], \qquad (1)$$

где $\psi_k[n]$ – выбранный на k-м уровне разложения материнский вейвлет, $j = 1, \dots, N_k$ – номер фильтра, $a = 2^{-j/Q_k}$ – коэффициент масштабирования, T – период выборки, n – индекс элемента выборки, Q_k обозначает количество фильтров на октаву в банке фильтров на k-ом уровне разложения.

Коэффициенты вейвлет-преобразования рассчитывались путем взятия абсолютных значений выходов N_k фильтров для соответствующего банка с заданной константой Q_k:

$$p_{kj}[n] = \left|\left(\psi_{kj} * x\right)[n]\right|, \qquad (2)$$

где x обозначал входной сигнал данного уровня разложения, * – операцию свертки, а |.| –оператор модуля для вычисления абсолютного значения.

Коэффициенты рассеяния на каждом уровне разложения получали путем кадрирования $p_j[n]$ на блоки равной длины M с использованием усредняющего фильтра (фильтра низких частот) $\phi[n]$, применяемого к каждому блоку:

$$S_k^{(j)}[m] = \left(\phi * p_{kj}\right)[mM], \qquad (3)$$

где $S_k^{(j)}[m]$ обозначает коэффициент разложения, соответствующий j-ому полосовому фильтру $\psi_{kj}[n]$ на k-ом уровне разложения, а m обозначает индекс блока.

Видно, что квадрат коэффициента рассеяния нулевого уровня является приближением кратковременной энергии речевого сигнала s[n], а квадраты коэффициентов рассеяния первого уровня являются оценками энергии в поддиапазонах, определенных набором фильтров постоянной Q_1 первого уровня. Разложения рассеяния второго уровня можно рассматривать как спектральное разложение мгновенных амплитуд компонент спектра s[n]. Другими словами, коэффициенты рассеяния второго уровня $S_2[m]$ являются «коэффициентами модуляции» спектра s[n].

Вычисленные коэффициенты рассеяния сигнала нескольких уровней использовались в качестве входных векторов в схеме вычисления векторов идентичности дикторов. Последние являются одним из стандартных представлений характеристик диктора и обычно представляют собой вектора размерности 103-104, которые характеризует уникальные особенности голоса, не зависящие от вносимых в

него искажений, связанных с процедурой фиксации голоса, шумами и другими искажениями. Вычисление векторов идентичности до середины 2010-х гг проводилось с помощью метода, основанного на адаптации базовых моделей голоса [5]. В настоящее время, однако, большее распространение получили нейросетевые методы извлечения данных представлений.

Архитектура глубокой нейронной сети, извлекающей низкоразмерные представления голоса [6], представлена на рисунок 2. Данная архитектура относится к так называемым нейронным сетям с временной задержкой (time-delay neural networks, TDNN-сетям). На вход сети подавались вектора признаков для заданного временного окна контекста. При обучении сети распространение ошибки происходило не только по ней непосредственно, но и по ее копиям, «сдвинутым» по временной шкале, что позволяло добиться инвариантности к временным сдвигам входного сигнала.

В таблице 1 приведены характеристики слоев TDNN-сети. Под фреймами 1-5 понимались слои с временными окнами, индивидуальный контекст и накопленная ширина контекста которых представлены во втором и третьем столбцах. Слой статистического пулинга вычислял по набор из двух статистик по всем входным данным сигнала – средние и среднеквадратичные отклонения выходов слоя «фрейм 5». Полученное представление размером 3000 подавалось на два полносвязных слоя – первый, сжимающий его до 512-мерного вектора значений и второй – «классифицирующий». Слой softmax выдавал решение о классе (идентичности субъекта, чей голосовой сигнал подан на вход сети). В качестве низкоразмерных признаков использовались выходные значения первого полносвязного слоя.

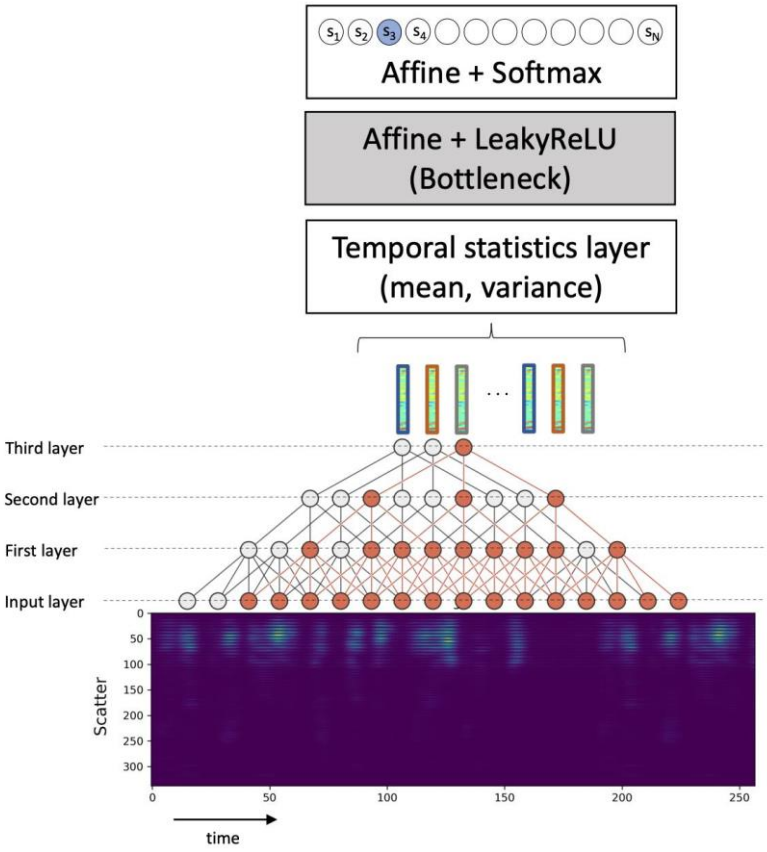

Рисунок 2. Архитектура применяемой TDNN-сети. Для упрощения в схеме приведена часть слоев с временными задержками.

В данной работе использовались образцы записей голоса из двух наборов данных VoxCeleb1 [7] и VoxCeleb2 [8]. Выбор этих наборов обуславливался

несколькими факторами. Данные наборы достаточно представительны и содержат образцы голоса более чем 7000 дикторов, в них представлено более 1 миллиона образцов записей (порядка 2000 часов звука. Эти наборы свободно распространяются для академических и исследовательских целей.

Таблица 1. Архитектура нейронной сети для извлечения низкоразмерных признаков аудиосигналов.

Слой	Временной контекст слоя (в номерах блоков)	Общий контекст (в номерах блоков)
Фрейм 1	[-2, +2]	[-2, 2]
Фрейм 2	[-1, 2]	{-1, 2}
Фрейм 3	[-3, 3]	{-3, 3}
Фрейм 4	[-7, 2]	{-7, 2}
Фрейм 5	{0}	{0}
Статистический пулинг	[0, T)	весь сигнал
Полносвязный слой 1	{0}	весь сигнал
Полносвязный слой 2	{0}	весь сигнал
Softmax	{0}	весь сигнал

Набор данных VoxCeleb1 [7] содержал более 100000 аудиозаписей от 1251 диктора. Они были получены из видеороликов социальной сети YouTube и представляют собой голоса известных персон (селебрити). Набор данных гендерно сбалансирован (55% мужчин и 45% женщин). Представлен широкий набор этнических групп, акцентов, профессий и возрастов. Запись проводилась в широком спектре окружений – от открытых стадионов до тихих комнат и студий звукозаписи. Первоначальным предназначением этого набора данных было проведение соревнования по машинному обучению по идентификации/верификации дикторов как на основе аудиоинформации, так и мультимодально на основе видео.

Кратко состав набора данных VoxCeleb1 представлен в таблице 2. Он был разделен на два подмножества: Development и Test. Первый предполагалось использовать для обучения системы верификации или идентификации, второй – для тестирования обученной системы.

Таблица 2. Структура набора данных VoxCeleb1.

	Development	Test
Разделение образцов для задачи верификации		
Число дикторов	1211	40
Число записей	148642	4874
Разделение образцов для задачи идентификации		
Число дикторов	1251	1251
Число записей	145265	8251

Второй набор данных VoxCeleb2 [8], представлял собой расширение первоначального набора данных, предназначенное в первую очередь для решения задачи обучения сложных глубоких нейросетевых моделей, которым для сходимости требуются большие объемы обучающих данных. Аналогично первому набору

данных в VoxCeleb2 были представлены аудиозаписи из YouTube, сделанные в широком диапазоне условий, для широкого круга дикторов, с достаточно представительным гендерным составом (61% мужчин, 39% женщин), в различных шумовых окружениях. Аналогично первому набору данных, второй был разделен на обучающее и тестовое подмножества (подробнее – в таблице 3).

Таблица 3. Структура набора данных VoxCeleb2.

	Development	Test
Число дикторов	5994	118
Число записей	1092009	36237

Все аудиозаписи предварительно были приведены к одинаковой длине – 5 секундам (80000 отсчетов) при частоте дискретизации 16кГц. Если аудиозапись оказывалась короче 5 секунд, то она увеличивалась путем добавления повторов этого же аудио пока общая длина не достигнет необходимого. Если же речевой сигнал был длиннее 5 с, то из него выбирался случайный отрезок размером в 80000 отсчетов. Далее для полученного набора данных было проведено предварительное акцентирование [9] с коэффициентом 0.97. Для полученного набора данных вычислялись коэффициенты преобразования рассеяния при $Q_k = 12$ и Nk=8. Использовалось три уровня разложения (k=0,1,2), которые рассматривались как одномерный вектор, характеризующий каждый блок сигнала. В качестве функций материнского вейвлета и функции усреднения по блоку на всех уровнях использовались вещественная часть вейвлета Морле и гауссова функция соответственно [10].

Нейронная сеть для извлечения низкоразмерных представлений обучалась при помощи оптимизатора ADAM [11] (с параметрами learning rate = 10-3, wd = 0). В качестве слоя принятия решения использовался стандартный softmax. Функцией потерь была кросс-энтропия. Все функции активации сети в отличие от [6] были заменены на LeakyRelu [12] версии.

При верификации дикторов проводилось сравнение схожести извлекаемых сетью 512-мерных векторов признаков. Для этого применялась косинусная мера сходства:

$$\text{similariy}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}, \quad (4)$$

где A и B – это некоторые низкоразмерные вектора признаков двух дикторов. На основе данной метрики вычислялись значения порогов принятия решения для обучающих выборок и производилась оценка качества работы системы верификации на выбранных тестовых образцах. В качестве метрики для сравнения применялась EER [1]. Результаты численных экспериментов приведены в таблице 4.

Таблица 4. Результаты апробации разработанного подхода.

Датасет обучения	Датасет тестирования	EER, %
Voxceleb2	Voxceleb2	**8,64**
Voxceleb2	Voxceleb1	9,18
Voxceleb2 (мужчины)	Voxceleb1	14,42
Voxceleb2 (женщины)	Voxceleb1	16,61
Voxceleb1 (мужчины)	Voxceleb2 (мужчины)	9,32
Voxceleb1 (женщины)	Voxceleb2 (женщины)	10,98

Видно, что использование при обучении большой представительной базы образцов (всего обучающего набора Voxceleb2), значительно снижает ошибку принятия решения. Полозависимые модели ожидаемо дали снижение качества при проверке на полных наборах данных. Применение модели, обученной на одном наборе данных, для данных другого набора продемонтстрировало слабое снижение качества (с 8,64% до 9,18%). Таким образом, можно предварительно говорить о возможном преимуществе предлагаемого подхода в тех ситуациях, когда доступ к непосредственным данным зарегистрированных пользователей ограничен, а имеется возможность обучения нейросетевой модели только на независимом наборе речевых данных.

Библиографический список

1. Rabiner L., Juang B.H. Fundamentals of speech recognition // N.-J. PrenticeHall, 1993. – 507 p.
2. ГОСТ Р 58624.1–2019. Информационные технологии. Биометрия. Обнаружение атаки на биометрическое предъявление. Стандарт по атакам представлением. Часть 1. Структура
3. Mallat S. Group Invariant Scattering [электронный ресурс] // режим доступа: http://arxiv.org/abs/1101.2286.
4. Anden J., Mallat S. Multiscale Scattering for Audio Classification // Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011. pp. 657-662.
5. Verma P, Das PK. I-vectors in speech processing applications: a survey // International Journal of Speech Technolng. — 2015. — Vol. 18, No. 4. DOI: 10.1007/978-981-10-6626-9_18.
6. Snyder D., Garcia-Romero D., Sell G., Povey, D., Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition // ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). –pp. 5329-5333.
7. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large scale speaker identification dataset [электронный ресурс] // режим доступа: https://arxiv.org/pdf/1706.08612
8. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition [электронный ресурс] // режим доступа: https://arxiv.org/pdf/1806.05622
9. Huang X., Acero A., Hon H.-W. Spoken Language Processing. A Guide to Theory Algorithm and System Development. N.-J. Prentice Hall. – 965 p.
10. Lee Fugal D. Conceptual Wavelets in Digital Signal Processing // San Diego: Space & Signals Technologies. 2009. 302 p.
11. Kingma D., Ba J. Adam: A Method for Stochastic Optimization // Proc. of International Conference on Learning Representations [электронный ресурс] // режим доступа:: https://arxiv.org/pdf/1412.6980
12. Pedamonti, D. Comparison of non-linear activation functions for deep neural networks on MNIST classification task [электронный ресурс] // режим доступа:: https://arxiv.org/pdf/1804.02763

USE OF SCATTERING TRANSFORM ON DISCRETE WAVELET DECOMPOSITION COEFFICIENTS FOR BIOMETRIC SPEAKER VERIFICATION

Lependin A.A., Gaponov D.A., Filin Y.A., Ladygin P.S.
Altai State *University, Barnaul*
email: andrey.lependin@gmail.com

Abstract. In this paper authors propose a new approach for calculating of speech signal features for the sake of speaker verification problem. A multilevel transformation was applied to the signal, calculating the scattering coefficients based on discrete wavelet decomposition. The resulting feature vectors were used as input data for a time-delay neural network. On their basis, the neural network calculated the speaker identity vectors, which were directly used for biometric verification. The proposed approach was tested on data from the VoxCeleb1 and VoxCeleb2 voice sample sets. The effectiveness of the approach was shown in comparison with existing verification methods based on deep neural networks.

Keywords: voice verification, discrete wavelet transform, scattering transform, time-delay neural network, speaker identity vector