АНАЛИЗ ПРИЗНАКОВ ИЗ WAVENET АВТОЭНКОДЕРА В ЗАДАЧЕ ОБНАРУЖЕНИЯ ИСКУССТВЕННЫХ ИСКАЖЕНИЙ В АУДИОФАЙЛАХ

И.Д. Ильяшенко, Р.С. Насретдинов
Алтайский государственный университет, г. Барнаул

В связи с развитием технологий высококачественной записи и редактирования аудиосигналов проведение фоноскопической экспертизы при анализе подлинности аудиозаписей наталкивается на существенные трудности. Эксперт должен учитывать возможность того, что потенциальный злоумышленник может внести существенные изменения в аудиозапись, незаметные как на слух, так и при использовании существующих методов экспертизы.

Задача определения подлинности и неизменности записи при фоноскопическом исследовании как правило решается при помощи анализа большого числа признаков: сдвигов контуров аудио фреймов, наличия следов применения антиалиасингового фильтра, изменения спектральных характеристик записи и/или фоновых шумов, появления безосновательных пауз в записанной беседе, изменений естественного энергетического уровня записанного сигнала и акустической глубины, искажений параметров цифровой сигналограммы, неожиданных скачков постоянной компоненты аудиосигнала. Как правило, анализ проводится с использованием множества специализированных программных утилит и предполагает большой практический опыт у эксперта.

В [9] был предложен подход к обнаружению изменений, вносимых в аудиосигнал, позволяющий проводить анализ аудиозаписей в полуавтоматическом режиме. Он может позволить существенно ускорить работу эксперта-фоноскописта. В той же работе был создан датасет для обучения предложенной модели, в котором искажения в аудиофайл вносятся в автоматическом режиме. Очевидно, что эти данные далеки от реальных, где искажения производятся вручную.

В данной работе этот датасет был проанализирован и сравнен с аудиофайлами, «склейки» в которых производились профессиональным музыкантом с наложением фреймов.

В основе подхода лежит предположение о том, что для классификации информации, содержащейся в аудиофайлах можно эффективно применять вектора признаков, извлекаемые из нейросетевой генеративной модели. В данной работе в качестве таковой был выбран автокодировщик (автоэнкодер) на основе

нейросетевой архитектуры WaveNet [1], исходно спроектированной для синтеза музыкальных и речеподобных аудиосигналов. Он может автоматически обучаться представлению данных на входе, восстанавливая их на выходном слое. При этом в нейронной сети имеется выделенный внутренний слой, имеющий значительно меньшую размерность по сравнению с входной или выходной [2]. Внутренний слой представляет собой некоторую «сжатую» форму представления аудиосигнала, обрабатываемого сетью. Полученное «сжатое» представление и будет использоваться для анализа.

Для обучения классификаторов и апробации предложенного подхода была создана сбалансированная база данных искусственно искажённых аудиозаписей на основе образцов из речевого англоязычного корпуса CSTR VCTK [3], который состоял из записей голосов 109 носителей английского языка, каждый из которых произносил по 400 предложений. Основу произнесённых текстов составляли газетные статьи.

При внесении искусственных искажений выбирались равномерно расположенные в каждом аудиообразце моменты времени с шагом 512 отсчетов (длительностью 32мс при частоте дискретизации 16кГц), начиная с 256-го отсчета. Каждый второй из этих моментов «склеивался» со значениями аудиосигнала, находящимися через 8000 отсчетов (через 0,5 с) после него (рис. 1). Размер каждого обучающего примера соответственно был равным T=512 отсчетам, что соотвествовало размеру воспринимающего поля предобученного WaveNet-автокодировщика, описанного выше.

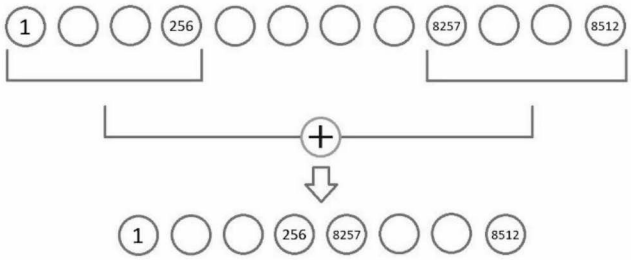

Рис. 1. Внесение искусственных искажений в аудиофреймы.

Для созданных образцов искаженных и неискаженных аудиосигналов с помощью автокодировщика были вычислены вектора признаков, которые в последствии будут называться «Выборка А».

Также для обучения модели были взяты профессионально обработанные образцы склейки. Склейка производилась путём наложения различных сигналов место склейки и размер наложения подбирались в промежутке между произносимыми словами с условием наложения амплитуд звукового сигнала.

Для анализа искажённого сигнала брались 512 отчётов полностью включающие фрагмент, содержащий наложение. В качестве неискажённых образцов выбирался фрагмент в паузе, не содержащий искажений. На этих аудиосигналов также были вычислены вектора признаков с помощью автокодировщика, которые впоследствии будут называться «Выборка B».

Для визуализации выборки A был применён алгоритм t-SNE [4]. Алгоритм не смог кластеризовать данные, однако были выявлены две интересные особенности. Во-первых, линейная направленность данных, что видно из рисунка 2. Во-вторых, была замечена принадлежность общему пространству без относительно используемых меток.

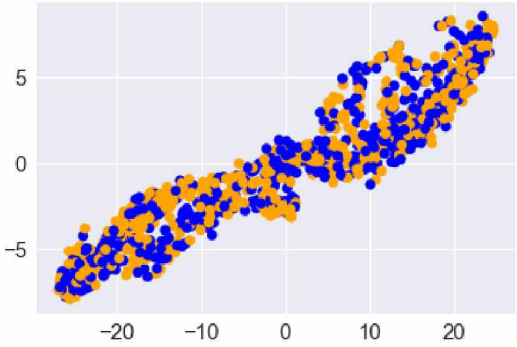

Рис. 2. Результат работы алгоритма t-SNE на выборке A.

Для выборки A была построена матрица корреляции признаков [5], которая представлена на рисунке 4. Из представленной матрицы видно, что полученные признаки имеют высокую корреляцию. 7 признаков имеют степень корреляции более 0,95. Было принято решение проводить дальнейший анализ без использования данный признаков. В дальнейшем анализе будут учувствовать признаки {f0, f2, f5, f6, f7, f9, f10, f12, f15}.

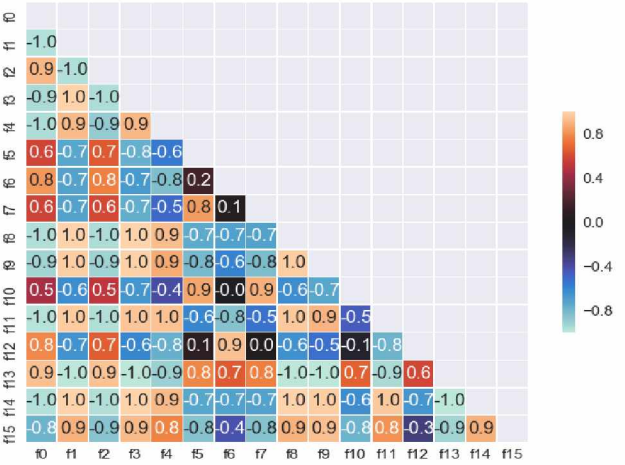

Рис. 4. Корреляция признаков выборки A.

Для каждого признака были построены диаграмма размаха (box plot) [6] и скрипичный график (violin plot) [7]. Графики всех признаков, полученных из автокодировщика, схожи с представленными на рисунке 5. Признаки различных классов имеют бимодальное распределение, но их распределения являются явно визуально различимыми. В то же время медианные значения признаков для каждого класса не выходят за квантили 25,75 для всех признаков, что сходится с замеченным отсутствием кластеризации.

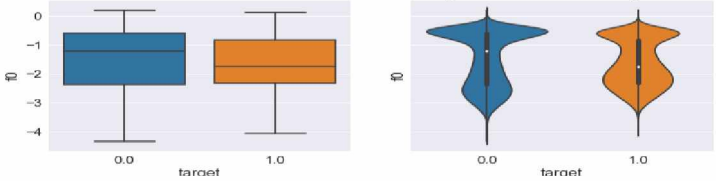

Рис. 5. Диаграмма размаха и скрипичный график параметра f_0 выборки A.

Также на выборке A были просчитаны PCA компоненты [8]. PCA или метод главных компонент на сегодняшний день является одним из основных способов уменьшения размерностей данных. Нахождение главных компонент сводится к вычислению сингулярного разложения матрицы данных из которого мы можем

получить ортогональные проекций с наибольшим рассеянием, которые и будут является искомыми компонентами.

На рисунке 6 представлено соотношение процента описанного разнообразия в данных от количества используемых PCA компонент. Первая PCA компонента уже описывает почти 95% разнообразия данных, что соотносится с полученными ранее результатами. Для проверки достаточности 95% разнообразия данных при классификации было построено дерево принятия решений. Параметры дерево подбирались с помощью кросс-валидации. Полученный при обучении результат на тестовой выборке был равен 0,65 по метрике roc_auc, что, по сравнению с полученными ранее результатами 0,9, свидетельствует о недостаточности содержащейся в одной PCA компоненте данных для классификации, несмотря на столь большой процент описываемого разнообразия данных.

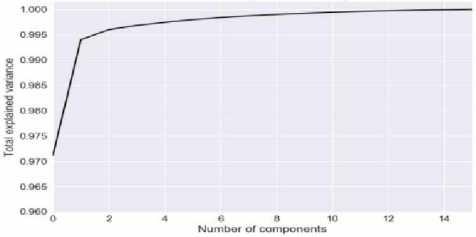

Рис. 6. Описание разнообразия данных PCA компонентами в выборке А.

Для выборки Б при использовании метода уменьшения размерности t-SNE [4] данные разделились на небольшие кластеры, что показано на рисунке 7. На верхнем графике можно выделить участок в левой части, состоящий в основном из оранжевых точек, и участок в правом верхнем углу и внизу– состоящие из синих. Но на остальной части графика искаженные и неискаженные примеры смешиваются.

Для выборки Б также были построена матрица корреляции, которая представлена на рисунке 8. Из рисунка видно, что признаки также сильно коррелируют между собой. Для упрощения дальнейшего анализа признаки с коэффициентом корреляции более 0,95 были отброшены.

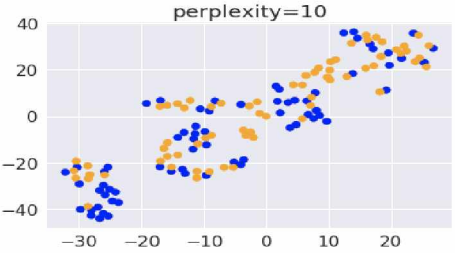

Рис. 7. Результат работы алгоритма t-SNE на выборке А.

Рис. 8. Матрица корреляция признаков выборки Б.

На рисунке 9 представлены диаграмма размаха и скрипичный график. Медианные значения на выборке Б в среднем отличаются слабее, чем на предыдущей, но на некоторых признаках видны явные отличия. Для параметров f0, f2, f6 медианные значения, а также значения первой и третей квантили, заметно различаются для искаженных и неискаженных участков.

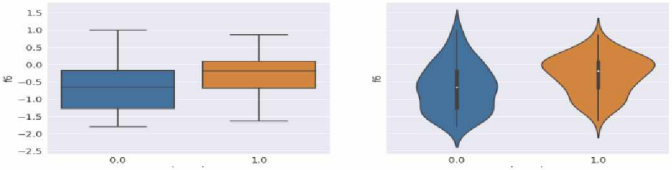

Рисунок 9. Диаграмма размаха и скрипичный графики параметра f_6 выборки Б.

Для проверки применимости выборки А в качестве обучающей для классификации на выборке Б. На выборке А был обучен XGBoost классификатор [9]. Подбор параметров XGBoost классификатора проводился путём кросс-валидации методом k-fold с разбиением обучающей выборки на 5 частей. Значения метрики roc_auc и F1-мера на отложенной выборке для подобранной модели равны 0,9 и 0,9 соответственно. К данным из выборки Б был применён представленные выше обученный алгоритм классификации. В таблице 1 приведены полученные результаты. Видно, что полученное очень низкое качество классификации, что говорит о кардинальных различиях в данных и невозможности применения моделей на из других данных.

Таблица 1. Результат классификации выборки Б

Классификатор	Точность	Полнота	F1-мера	Roc auc
XGBoost	0.85	0.86	0.85	0.85

Для оценки важности параметров с коэффициентом корреляции более 0,95 в выборки А для поставленной задачи определения подлинности было проведено обучение XGBoost классификатора с использованием параметров {f0, f2, f5, f6, f7, f9, f10, f12, f15}. Подбор оптимальных параметров проводился методом k-fold с разбиением на 5 частей. Для сравнения качества предсказаний были выбраны следующие метрики: точность, полнота и F1-мера, roc auc. Из результатов, представленных в таблице 2, видно, что все метрики упали на значение 0,04, что свидетельствует о малом содержании полезной для классификации информации в отброшенных признаках, но не их ненужности. Признаки с высокой степенью корреляции вносят небольшой вклад, но всё же позволяют улучшить результат классификации.

Таблица 2. Результат обучения выборки А

Классификатор	Точность	Полнота	F1-мера	Roc auc
XGBoost	0.85	0.86	0.85	0.85

В работе был проведен анализ двух выборок искусственно искаженных аудиофайлов: с автоматической «склейкой» участков и профессиональной, с наложением фреймов. Визуально данные сильно отличаются, но сходства все же присутствуют, что хорошо видно на матрицах корреляции и диаграммах распределения. При классификации данных из выборки с профессиональной «склейкой» было получено неудовлетворительное качество – 63% верных предсказаний против 90% на выборке с автоматической «склейкой».

Такое значение получилось ввиду присутствия отличий в данных для обучения и тестирования моделей, что было показано в ходе работы.

Анализ выборки с профессиональной «склейкой» показал, что разделение между искаженными и неискаженными аудиофайлами все же присутствует. Поэтому следующим этапом развития научно-исследовательской работы будет сбор большего количества профессионально искаженных данных, для последующего обучения.

Библиографический список

1. Guyon I., Dior G., Lemaire V. et al Autoencoders, Unsupervised Learning, and Deep Architectures [Электронный ресурс] / 2012. – Режим доступа: http://proceedings.mlr.press/v27/baldi12a/baldi12a.pdf
2. Engel J., Resnick C., Roberts A. et al Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders [Электронный ресурс] / 2012 - Режим доступа: https://arxiv.org/pdf/1704.01279.pdf
3. Veaux Ch., Yamagishi J., MacDonald K. CSTR VCTK Corpus. English Multi-speaker Corpus for CSTR Voice Cloning Toolkit [Электронный ресурс] / 2010. – Режим доступа: http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html
4. Laurens van der Maaten., Geoffrey Hinton._Visualizing Data using t-SNE [Электронный ресурс] / 2008. - Режим доступа: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
5. M. E. Tipping., C. M. Bishop. Mixtures of Probabilistic Principal Component Analysers. [Электронный ресурс] / 2006. - Режим доступа: http://www.miketipping.com/papers/met-mppca.pdf
6. Yoav Benjamini. Opening the Box of a Boxplot / Yoav Benjamini // The American Statistician. – 1988. – 4. - pp. 257-262
7. _Violin Plots 101: Visualizing Distribution and Probability Density/ Joel Carron [электронный ресурс] / 2016. - Режим доступа: https://blog.modeanalytics.com/violin-plot-examples/
8. M. Tipping, Probabilistic Principal Component Analysis / M. Tipping, C. Bishop // Journal of the Royal Statistical Society - серия B – 61 - часть 3 - с. 611-622. – Режим доступа: http://www.miketipping.com/papers/met-mppca.pdf
9. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System [Электронный ресурс] / 2016. – Режим доступа: https://arxiv.org/abs/1603.02754v3
10. И.Д. Ильяшенко., Р.С. Насретдинов., А.А. Лепендин. Применение wavenet-автоэнкодера в задаче обнаружения искусственных искажений аудиофреймов / 2018.