УДК 004.93’1

АВТОМАТИЧЕСКОЕ ОБНАРУЖЕНИЕ ЗВУКОВЫХ ДЕФЕКТОВ В АУДИОФАЙЛАХ

© 2019. *М.В. Мартынов, Т.В. Шарий*

В статье предлагается метод автоматического обнаружения звуковых дефектов в аудиосигналах, опирающийся на эффективные модели машинного обучения. Описана общая схема и алгоритм расчета вектора признаков звукового фрейма на основе амплитудно-временного, спектрального и кепстрального анализа сигнала. Приведены результаты экспериментов по детектированию аудиодефектов в музыкальных файлах с помощью многослойного персептрона и деревьев решений.

Ключевые слова: звуковой дефект; глитч; спектральные признаки; MFCC; машинное обучение.

Введение. В последнее время наблюдается тенденция к увеличению объемов аудиоколлекций в цифровом виде как для личного пользования, так и других всевозможных целей, благодаря таким форматам звука, как MP3, FLAC и OGG [1]. На сегодняшний день самым главным источником пополнения музыкальной фонотеки для большинства пользователей является интернет, где можно легко найти предпочтительный контент и загрузить его из множества ресурсов. Тем не менее, качество звука в некоторых случаях может оказаться недопустимо низким для слушателя. Звук может иметь искажения, такие как: глитч-эффекты [2], скачки, плохое выравнивание (эквализация), клиппирование, шум, неполнота, растяжение / сжатие во времени [3]. Аудиоданные могут быть заявлены как оригинальные, однако в действительности являться восстановленными после сжатия с потерями. Кроме того, при копировании данных с компакт-дисков не исключены ошибки чтения из-за механических проблем привода или некачественной поверхности диска. Динамический диапазон закодированных треков также может варьироваться в значительных пределах, что в конечном итоге повлияет на качество итогового файла.

Важную роль играют и представляют интерес системы цифрового восстановления аудиоматериалов, накопленных в большом количестве к текущему времени за эпоху кассет и VHS-видео, а также звука с виниловых пластинок, несмотря на заметную тенденцию возвращения меломанов к данному виду звуковых носителей и проигрывателей.

Таким образом, становится весьма актуальной задача автоматического обнаружения перечисленных аудиодефектов. Дальнейшие действия уже могут изменяться, в зависимости от ситуации: удаление некачественного аудиоматериала либо попытка исправления дефекта. Проблемы такого рода на протяжении многих лет успешно решались с помощью методов цифровой обработки сигналов [4,5], а в последнее время к ним подключаются мощные современные модели машинного обучения [6-9], что позволяет достигать значительных результатов.

Постановка задачи. Целью работы является повышение качества автоматического обнаружения звуковых дефектов в аудиофайле. В статье не рассматриваются дальнейшие действия по удалению дефектов или отбраковке некачественного аудиоматериала. Информационная технология обнаружения аудиодефектов должна поддерживать процессы загрузки и визуализации звуковых файлов, фильтрации, параметризации и аугментации звукового сигнала, сохранения

параметров в файл, обучения классификаторов на основе вычисленных параметров, визуализации результатов обучения и распознавания.

Общая схема автоматического обнаружения аудиодефектов. Любой звуковой дефект является результатом повреждения источника аудиоданных. Это произвольное искажение сигнала, возникающее по причине ошибок при выполнении одной из следующих процедур: 1) оцифровка (характерный шум виниловой пластинки, поврежденный винил, ошибки CD-проигрывателя, царапины на компакт-диске, износ пленки кассеты); 2) кодирование / декодирование (ошибки в алгоритмах кодирования); 3) сведение (мастеринг) (низкое качество используемого оборудования, ошибки стереомикширования); 4) запись в плохих условиях (шумная окружающая обстановка, дефекты оборудования); 5) сжатие в более экономный формат, например, MP3; 6) цифровая фильтрация (нежелательное подавление или поднятие определенных частот).

Одним из наиболее распространенных аудиодефектов является т.н. глитч. Обнаружение глитча в сигнале сводится к оконному анализу сигнала и в каждом окне к задаче бинарной классификации «Есть дефект / нет дефекта».

Общая схема работы системы автоматического обнаружения звуковых дефектов в аудиофайлах, изображенная на рис.1, представляет собой классический двухуровневый вариант решения задачи машинного обучения вида «расчет признаков – обучение и распознавание на основе признаков».

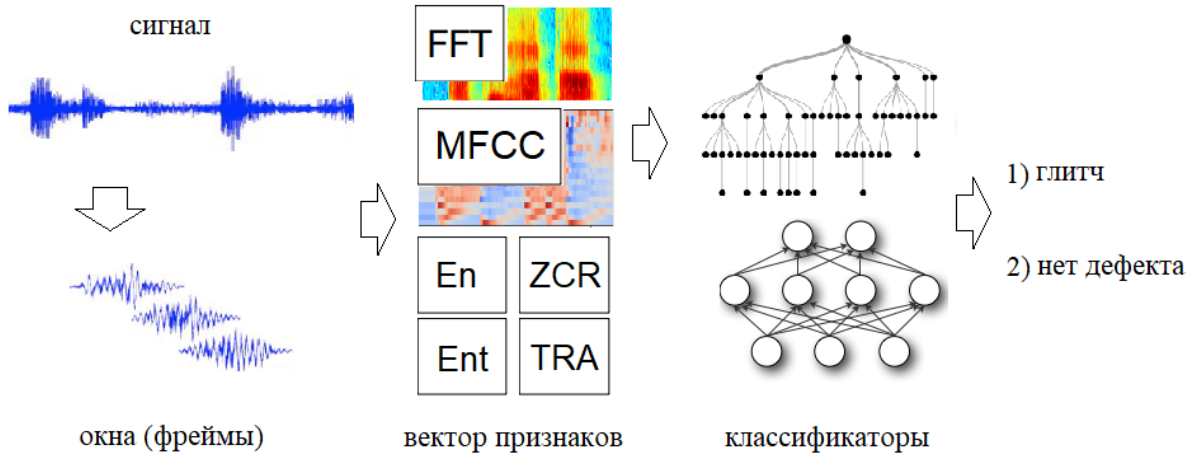

Рисунок 1 – Общая схема обнаружения звуковых дефектов в аудиофайле

Несмотря на то что, на втором этапе применяются модели глубокого обучения, в статье вопросу инженерии признаков (их извлечению и отбору) уделяется главное внимание. Предполагается, что в конкретном аудиофрагменте присутствует не более одного звукового дефекта. Процедура извлечения всех релевантных признаков звука производится в окнах длительности 100 мс с перекрытием 25 мс, в которых каждый взвешенный оконной функцией Хемминга фрейм становится источником вектора признаков. На основе уровнеграммы аудиосигнала вычисляется ряд полезных амплитудно-временных характеристик звука, которые добавляются в вектор признаков, наряду со своими статистиками, описываемыми в следующем подразделе работы. На основе спектра Фурье также рассчитывается ряд характеристик, входящих в вектор признаков аудиофрагмента. Также на основе спектра сигнала производятся дальнейшие действия для мел-частотного кепстрального анализа с последующим включением

кепстральных статистик в вектор принаков. Данными векторами оперируют статистические модели-классификаторы на этапе распознавания звуковых дефектов и на этапе обучения: нейронные сети и деревья принятия решений. Они сначала обучаются на большом наборе данных (подбирают веса на основе всех векторов признаков из обучающей выборки аудиодефектов), после чего применяются для вычисления типа аудиодефекта (или его отсутствия) в текущем аудиофрагменте. Для хранения всех вспомогательных и промежуточных данных используются csv-файлы.

Музыкальный сигнал загружается из аудиофайла любого из самых популярных на сегодняшний день форматов (MP3, OGG или WAV) либо записывается с микрофона и представляет собой дискретный набор отсчетов. Файлы формата MP3 и OGG предварительно декодируются в формат PCM (импульсно-кодовая модуляция (Pulse Code Modulation). Все сигналы приводятся к единой частоте дискретизации, равной 44,1 кГц; количество бит на отсчет – ко значению 16.

Параметризация звукового сигнала. Вектор признаков аудиосигнала включает 15 амплитудно-временных характеристик, 9 спектральных дескрипторов и 12 мел-частотных кепстральных коэффициентов.

Первая группа признаков вычисляется на основе уровнеграммы сигнала. Основными амплитудно-временными параметрами сигнала являются: 1) энергия участка; 2) энтропия участка сигнала; 3) частота перехода уровня сигнала через ноль; 4) длительность участка; 5) маркеры онсетов (начала и окончания звучания фрагмента); 6) различные статистики вышеуказанных параметров (среднее, дисперсия, стандартное отклонение, медиана, мода и т.д.).

Энергия участка аудиосигнала рассчитывается во временной области сигнала по простой формуле:

$$E(k,L) = \frac{1}{L}\sum_{i=k}^{k+L-1} x_i^{\,2},$$

где (и в дальнейших формулах) k – позиция начального отсчета участка; L – число отсчетов на анализируемом участке; x_i – значение i-го отсчета сигнала.

Энтропия участка аудиосигнала рассчитывается на основе равномерно распределенных бинов всех значений отчетов сигнала по формуле:

$$H(k,L) = -\sum_{i=1}^{nbins} P_i(x_k)\log_2 P_i(x_k),$$

где *nbins* – количество бинов распределения; $P_i(x_k)$ – вероятность принадлежности отсчета x_k *i*-ому бину.

Еще одним параметром речевого сигнала, который рассчитывается в статье на основе амплитуд его отсчетов, является частота перехода уровня сигнала через ноль (Zero Crossings Rate, ZCR) [5]. Эта характеристика вычисляется по формуле:

$$ZCR(k,L) = \frac{1}{L}\sum_{i=k}^{k+L-1} \frac{\left|sign(x_{i+1}) - sign(x_i)\right|}{2},$$

Признак «Сумма абсолютных изменений» (ASC, Absolute Sum of Changes) [10] вычисляется по формуле:

$$ASC(k,L) = \sum_{i=k+1}^{k+L-1} | x_i - x_{i-1} | . \quad (1)$$

Индекс сложности аудиофрагмента (CE, Complexity Estimate) – оценка сложности временного ряда; более «сложные» ряды характеризуются большим количеством пиков, провалов и т.д. Данный признак вычисляется по формуле:

$$CE(k,L) = \sum_{i=k+1}^{k+L-1} (x_i - x_{i-1})^2 . \quad (2)$$

Признак «Асимметрия временного обращения» (TRA, Time Reversal Asymmetry) вычисляется по формуле:

$$TRA = \mathrm{E}[L^2(X)^2 \cdot L(X) - L(X) \cdot X^2], \quad (3)$$

где *L(X)* – оператор задержки, введенный в [15], E – математическое ожидание. Формулу (3) можно развернуть в более подходящем для расчетов виде:

$$TRA = \frac{1}{L - 2lag} \sum_{i=0}^{L-2lag} x_{i+2lag}^2 \cdot x_{i+lag} - x_{i+lag} \cdot x_i^2 ,$$

где *lag* – значение задержки (количество семплов задержки); *L* – число отсчетов на анализируемом участке.

Помимо перечисленных характеристик, интерес представляют такие статистики, как: медиана, среднее, количество пиков, положение первого и последнего пиков, количество провалов, положение первого и последнего провалов. Данные параметры также включаются в вектор признаков.

Далее вектор признаков расширяется *спектральными параметрами,* которые вычисляются на основе спектра Фурье фрейма.

Спектральная равномерность (SFM, Spectral Flatness Measure) характеризует степень тональности или шумности сигнала в данном частотном диапазоне:

$$SFM = \frac{\sqrt[h-l+1]{\prod_{m=l}^{h} |f(m)|^2}}{\frac{1}{h-l+1} \sum_{m=l}^{h} |f(m)|^2} . \quad (4)$$

В формуле (4) и дальнейших формулах применяются следующие обозначения: *l* – левая (нижняя) граница анализируемого частотного диапазона спектра текущего фрейма (окна анализа) сигнала; *h* – правая (верхняя) граница анализируемого частотного диапазона спектра фрейма сигнала; *f(m)* – зависимость амплитуды от частоты (амплитудный спектр).

Спектральный центроид (SC, Spectral Centroid) определяет центр масс и выражает степень «яркости» звучания сигнала:

$$SC = \frac{\sum_{m=l}^{h} m\left|f(m)\right|^2}{\sum_{m=l}^{h}\left|f(m)\right|^2} \quad .$$

Спектральная энтропия сигнала (SE, Spectral Entropy) вычисляется по формуле:

$$SE = \sum_{m=l}^{h}\left|f(m)\right|\log_2\left|f(m)\right| .$$

Частота спада спектральной энергии (Spectral Roll-off) – это частота *Rf*, ниже которой содержится 85% (β=0.85) суммарной энергии спектра:

$$\sum_{m=1}^{Rf}|\ f(m)\ | = \beta \cdot \sum_{m=1}^{fs/2}|\ f(m)\ |,$$

где *fs* – частота дискретизации аудиосигнала.

Кепстральные параметры фрейма рассчитываются по алгоритму MFCC (Mel-frequency cepstral coefficients, мел-частотные кепстральные коэффициенты) [4]. В данном алгоритме последовательно выполняются следующие блоки вычислений. В первом блоке вычисляются коэффициенты амплитудного спектра Фурье. Во втором блоке на вычисленный спектр накладывается набор из 20 перекрывающихся фильтров психоакустической шкалы мел в диапазоне частот от 100 Гц до 4000 Гц (каждый фильтр имеет частотную характеристику треугольной формы). В третьем блоке выполняется логарифмирование измененного спектра, после чего, в финальном блоке, производится дискретное косинусное преобразование второго типа.

Таким образом, вектор признаков аудиофрейма завершается 12 коэффициентами MFCC, рассчитанными по описанному алгоритму.

Описание эксперимента и анализ результатов. Т.к. в работе для решения задачи автоматического обнаружения звуковых дефектов применяются статистические классификаторы, то качеству набора обучающих и тестовых данных уделяется особое внимание. Музыкальные файлы для экспериментов выкачиваются из архива FreeMusicArchive (FMA) [11]. Можно также использовать персональную коллекцию музыки, однако данные FMA лучше сбалансированы по жанрам и звуковой структуре. Совокупное время звучания наименьшего набора файлов составляет 66 часов 40 минут.

Самые важные данные содержатся в наборе непосредственно звуковых дефектов, который выкачивается с сайта freesound.org. В корпусе содержатся 39 глитч-дефектов. Уровнеграммы и спектрограммы части из них проиллюстрированы на рис.2.

В качестве базовых классификаторов выбраны многослойный персептрон и деревья принятия решений. Персептрон содержит 3 слоя, включая 64 входных нейрона, 32 нейрона в скрытом слое и 1 нейрон в выходном слое. Скрытые слои имеют функцию активации ReLU, выходной слой – сигмоид.

Т.к. многослойный персептрон используется для классификации «Есть дефект / нет дефекта», в качестве функции потерь применяется бинарная кросс-энтропия. Оптимизация значений весов при обучении нейронной сети производится по адаптивному алгоритму RMSProp [7,8].

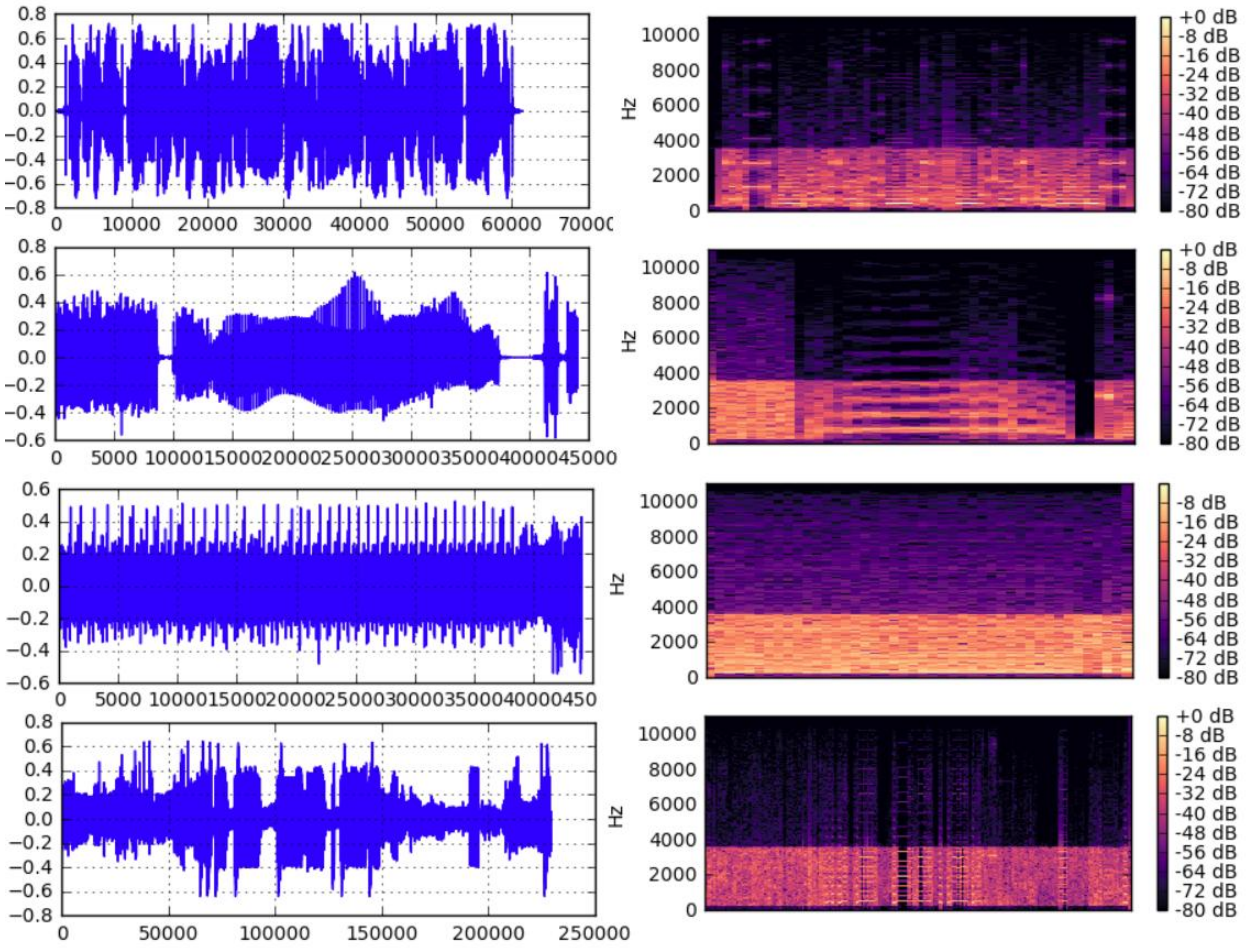

Рисунок 2 – Примеры уровнеграмм и спектрограмм глитч-дефектов
из репозитория dotY21 (freesound.org)

Эксперимент проводился по следующей методологии: 1) производилась аугментация данных и наложение каждого из обучающих глитч-эффектов в произвольных участках музыкальных сигналов; 2) в каждом из аудиосигналов выбирались по 5 произвольных участков и 5 искаженных, с сохранением позиций и меток; 3) в этих отобранных фрагментах, а также, частично, в сформированных пользователем фрагментах (в особо важных с точки зрения аудиоэксперта местах), вычислялись все необходимые амплитудно-временные, спектральные, кепстральные параметры для формирования вектора признаков; 4) формировались векторы признаков и сохранялись в csv-файле с метками «дефект» и «нет дефекта». В итоге получилось 9600 36-мерных записей (векторов признаков); 5) производилось обучение многослойного персептрона и дерева принятий решений с помощью скрипта audioml.py; 6) тестировались обученные модели на тестовой выборке, а также распознавались аудиодефекты во всем музыкальном сигнале с помощью скрипта recognize_defects.py.

Статистический анализ собранных данных показал, что часть признаков (коэффициентов) сильно коррелирует друг с другом. Также, как продемонстрировали первые эксперименты, данные в исходном виде вызвали большие трудности у классификаторов: точность распознавания деревьев принятия решений составляла

около 78-80% на тестовой выборке, в то время как многослойный персептрон даже на обучающей выборке не мог превысить базовый порог 50%.

В таблице 1 приведен список самых релевантных признаков, полученных в результате обучения ансамбля деревьев принятия решений на исходных 36-мерных векторах, отранжированных в порядке убывания их степени значимости.

Таблица 1 – Наиболее значимые дескрипторы аудиофрейма

Название дескриптора	Степень значимости
Сумма абсолютных изменений	0.192
Индекс сложности фрейма	0.153
Второй коэффициент MFCC	0.062
Третий коэффициент MFCC	0.057
Спектральный центроид	0.055
Шестой коэффициент MFCC	0.053

С учетом результатов разведочного анализа, данные подверглись дополнительной предварительной обработке: 1) нормализация к стандартному распределению (вычитание среднего и деление на стандартное отклонение); 2) сокращение размерности признакового пространства до 26 признаков с помощью анализа главных компонент (PCA, Principal component analysis). После обработки входных данных дерево принятия решений показало точность при кросс-валидации 80% (±2%); ансамблирование деревьев позволило повысить точность распознавания глитчей до 85%; многослойный персептрон на 60 эпохах показал, в среднем, точность 97% на обучающей и 91% на валидационных выборках (рис.3).

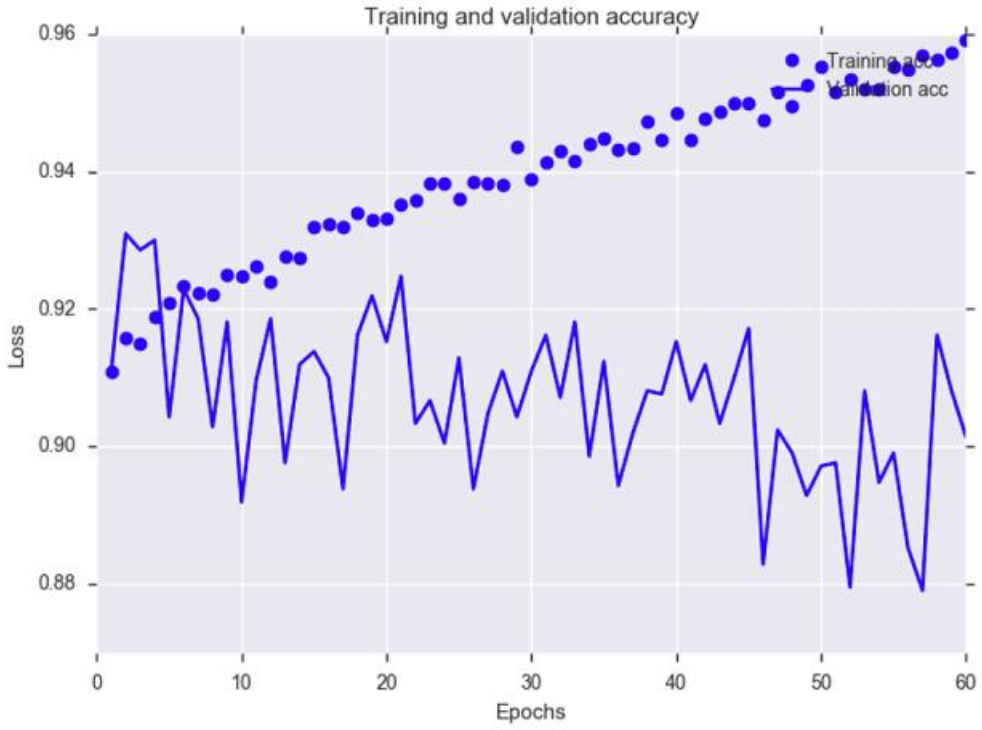

Рисунок 3 – Показатели работы многослойного персептрона

Этому результату можно доверять с дополнительными оговорками: размер выборки (как музыки, так и глитчей) был недостаточно велик для полноценного анализа эффективности моделей. Тем не менее, показатели свидетельствуют о том, что предложенный в статье подход перспективен.

Эксперименты производились на компьютере со следующими характеристиками: процессор Intel Core i3 с тактовой частотой 3,5 ГГц; объем оперативной памяти 16 Гб. Скорость расчета векторов признаков равна, в среднем, 85x (85 секунд сигнала обрабатываются за 1 секунду), что дает возможность использовать систему в реальном времени.

Выводы. Проведенные исследования продемонстрировали обоснованность и перспективность подхода к автоматическому обнаружению аудиодефектов в звуковых файлах на основе совместного использования амплитудно-временных, спектральных и кепстральных параметров сигнала. На относительно небольшой обучающей выборке музыкальных сигналов и глитчей с использованием простейших классификаторов машинного обучения получена точность бинарной классификации «дефект / не дефект» на отметке 91%. Корреляционный анализ значений признаков дал практические основания для сокращения размерности исходного признакового пространства, а анализ структуры дерева принятия решений показал, что среди всех параметров наиболее значимыми для классификации параметрами являются второй, третий и шестой коэффициенты MFCC, а также специальные дескрипторы «сумма абсолютных изменений» и «индекс сложности». Вычислительная сложность предложенного в статье метода относительно невелика и позволяет использовать его в реальном времени. Дальнейшая работа связана с более тщательным подбором и статистическим анализом низкоуровневых признаков сигнала, а также исследованием возможностей применения в задаче обнаружения аудиодефектов моделей глубокого машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

[1] Топ 10 аудиоформатов – рейтинг и преимущества [Электронный ресурс]. – Режим доступа: https://online-converting.ru/blog/audio-top10 / 02.03.2019.

[2] Reiss J. Audio Issues in MIR Evaluation / J. Reiss, M. Sandler // *Proceedings of the 5th International Conference on Music Information Retrieval ISMIR*. – 2004. – P.28-33.

[3] Godsill S. Digital audio restoration / S. Godsill, P. Rayner, O. Cappé // *Applications of digital signal processing to audio and acoustics*. – 2006. – P.133-194.

[4] Tralie C. Early MFCC and HPCP Fusion for Robust Cover Song Identification / C.Tralie // *Proceedings of the 18th ISMIR Conference, Suzhou, China*. – 2017. – P.294-301.

[5] Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project / G. Peeters // *CUIDADO Proj. Report*. – 2004.

[6] Choi K. A Tutorial on Deep Learning for Music Information Retrieval [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1709.04396v2.pdf / 02.03.2019.

[7] Николенко С. Глубокое обучение / С. Николенко, А. Кадурин, Е. Архангельская. – СПб.: Питер, 2018. – 480 с.

[8] Рашка С. Python и машинное обучение / С. Рашка. – М.: ДМК Пресс, 2017. – 418 с.

[9] Хайкин С. Нейронные сети: полный курс / С. Хайкин. – М.: Издательский дом «Вильямс», 2016. – 1104 с.

[10] Fulcher B.D. Highly Comparative Feature-based Time-series Classification / B. Fulcher, N. Jones // *IEEE Transactions on Knowledge and Data Engineering*. – Vol.26. – 2014. – P.3026-3037.

[11] Defferrard M. FMA: A Dataset for Audio Analysis / M. Defferrard, K. Benzi, P. Vandergheynst, X. Bresson // *Proceedings of the 18th ISMIR Conference, Suzhou, China*. – 2017. – P.316-323.

Поступила в редакцию 11.03.2019 г.

AUTOMATIC DETECTION OF SOUND DEFECTS IN AUDIO FILES

M.V. Martynov, T.V. Sharii

The article presents novel method for automatic detection of sound defects in audio signals relying on efficient machine learning models. The description is given of a general scheme and an algorithm for evaluating the features of a sound frame based on time-domain, spectral and cepstral analysis of signal. The results of experiments on audio defects detection in music files are given. In experiments multi-layer perceptrons and decision trees were used.

Keywords: *sound defect; glitch; spectral features; MFCC; machine learning.*