УДК 004.93

ХРОНОЛОГИЧЕСКОЕ УПОРЯДОЧЕНИЕ АУДИОФРАГМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ДВУХМЕРНЫХ СПЕКТРОГРАММ

А.Н. Алфимцев[1], **С.И. Назарова**[2]

[1]МГТУ им. Н.Э. Баумана, Москва, Российская Федерация
e-mail: alfim@bmstu.ru

[2]ОАО "Газпром автоматизация", Москва, Российская Федерация
e-mail: nazarova_svetlana92@mail.ru

Предложен метод анализа речевых аудиофрагментов и осуществления их хронологического упорядочения. Суть метода заключается в первоначальном представлении аудиофрагментов в виде двухмерных спектрограмм и затем анализа 1025 числовых дескрипторов, полученных как непосредственно из спектрограмм, так и из их преобразований. Значение сходства между двумя аудиофрагментами вычислено с использованием алгоритма K взвешенных ближайших соседей, по результатам работы которого построено дерево сходства для визуализации упорядочения речевых данных. В качестве материалов для эксперимента были взяты аудиофайлы — фрагменты выступлений известных политиков. Экспериментальное исследование подтвердило эффективность применения предлагаемого метода для хронологического упорядочивания аудиофрагментов, что с практической точки зрения открывает новые пути по разработке программных систем для автоматической обработки аудиоархивов и анализа характеристик речи.

***Ключевые слова*:** распознавание образов, двухмерная спектрограмма, вектор свойств, матрица сходства, хронологическое упорядочение.

CHRONOLOGICAL ORDERING OF THE AUDIO DATA USING 2D SPECTROGRAMS

A.N. Alfimtsev[1], **S.I. Nazarova**[2]

[1]Bauman Moscow State Technical University, Moscow, Russian Federation
e-mail: alfim@bmstu.ru

[2]OAO "Gazprom automation", Moscow, Russian Federation
e-mail: nazarova_svetlana92@mail.ru

The paper introduces an automatic quantitative method for both the speech fragments analysis and chronological ordering. The method consists of the following: audio fragments are initially presented in the form of two-dimensional spectrograms, then a large set of 1025 numerical descriptors extracting from both the raw spectrograms and their transforms is analyzed. The similarity value between two audio fragments is computed using a variation of the Weighted K-Nearest Neighbor scheme. A similarity tree is designed to visualize differences between the audio fragments. Some speech fragments of well-known politicians were used for the study. The proposed method proves to be efficient for chronological ordering of the audio fragments. It seems to introduce new ways of developing software systems for automated processing of audio archives and analysis of the speech characteristics.

***Keywords*:** pattern recognition, two-dimensional spectrogram, feature set, similarity matrix, chronological ordering.

Применение алгоритмов распознавания образов и машинного обучения к автоматическому анализу речевых отрезков позволяет решать множество задач. Одна из наиболее общих задач в области автоматического анализа аудиофрагментов — их классификация [1, 2]. В случае, например, с музыкальными произведениями классификация может быть осуществлена по жанрам [3, 4], эмоциональной окраске композиций [5], преобладающим музыкальным инструментам [6]. Другие направления исследований в области автоматического анализа музыки включают в себя автоматическую рекомендацию музыкальных произведений [7], поиск кавер-версий [8], предсказание качества звука [9] и др. Важной задачей также является предоставление возможности находить в базе музыкальных произведений наиболее похожие с исходным музыкальным отрезком [10, 11].

В настоящей статье впервые предложено применить метод анализа аудиофрагментов для автоматического анализа речевых отрезков (фрагментов выступлений). В области анализа речевых фрагментов могут ставиться такие задачи, как классификация данных по полу и возрасту говорящего (например, для проведения различных статистических исследований), по эмоциональной окраске речи, по времени записи и др. Наиболее сложной представляется задача расположения аудиофрагментов в хронологическом порядке. Даже человек (эксперт) далеко не всегда способен сделать это с высокой степенью точности. Автоматический анализ аудиозаписей позволяет отслеживать изменения в характеристиках речи человека на протяжении длительного периода времени. К этим характеристикам можно отнести паузальность, темп речи, силу, высоту и тон голоса.

Предложенный метод основан на детальном анализе аудиофрагментов, представленных своими двухмерными спектрограммами, а числовые дескрипторы (свойства) использованы для определения сходства между фрагментами, принадлежащими различным временны́м интервалам. Основное применение предложенного алгоритма — анализ речи в численных аспектах и для хронологического упорядочения данных (например, при автоматическом создании аудиоархивов), а также анализ и визуализация сходства в характеристиках речи при различных исследованиях (идентификация человека по голосу, обнаружение похожих голосов и т.д.).

Подготовка исходных данных. Для проведения точного анализа и проверки правильности работы метода необходимо иметь достаточное количество аудиоданных, принадлежащих одному человеку и записанных на протяжении нескольких лет. Исходя из этих требований, в качестве исходного набора данных были взяты фрагменты выступлений политиков, чья профессиональная деятельность началась не менее шести лет назад и материалы о которых (видео- и звукозаписи) есть в

открытом доступе. В проведенном исследовании были использованы записи, принадлежащие Бараку Обаме и Ангеле Меркель.

Исходные данные были разбиты на периоды длительностью в два года с момента начала деятельности политика на посту главы государства до настоящего времени. Компьютерный анализ, выполненный в настоящем исследовании, опирается на предположение, что два года — достаточный для анализа интервал, на котором у человека вырабатывается определенный стиль речи, выступления. Цель исследования — проверка возможности автоматического отслеживания этих изменений (распознавания динамики этих изменений) и хронологического упорядочения интервалов из входного набора данных.

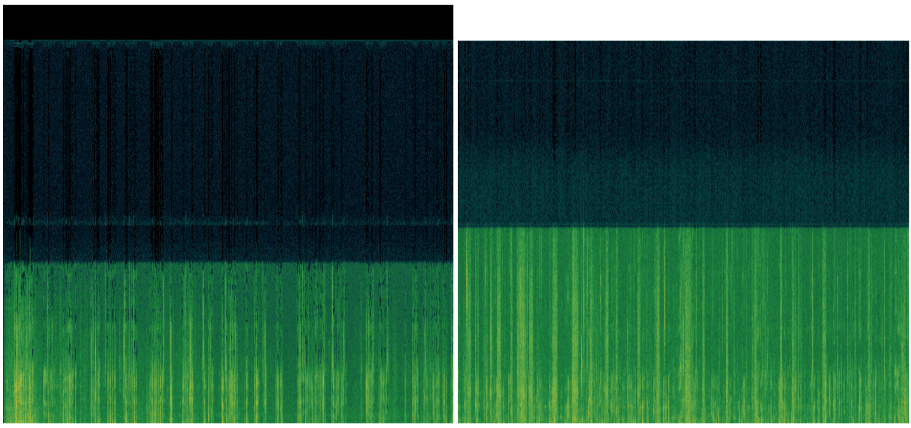
Каждый период включает в себя определенное количество аудиофрагментов, записанных в рассматриваемый интервал времени (для разных политиков это число варьируется от 13 до 18). По возможности было взято максимальное число отрезков из тех выступлений, где политик отвечает на вопросы журналистов, а не выступает с заранее подготовленной речью. Причина такого выбора исходных данных состоит в том, чтобы они максимально точно отражали характеристики речи человека в определенный период.

Аудиофрагменты первоначально были записаны в формате FLAC (Free Lossless Audio Codec), затем преобразованы в формат WAV (Waveform Audio File Format) моно. Для нормализации аудиоотрезков по длине из каждого аудиофайла был вырезан 60-секундный сегмент с помощью бесплатного он-лайн конвертера (www.online-convert.com). Эти фрагменты не содержат все выступление, но являются достаточно продолжительными для анализа характеристик речи. Аудиофайлы были выбраны так, чтобы в них не было посторонних шумов (разговоров, аплодисментов, помех аппаратуры и т.д.). Это было сделано для более объективного анализа данных.

Для проведения эксперимента была использована следующая разбивка входных данных: для анализа речи Барака Обамы взято 4 периода по 2 года каждый (2007–2009 гг., 2009–2011 гг., 2011–2013 гг., 2013 — по настоящее время). Для анализа речи Ангелы Меркель также взято 4 периода по 2 года каждый (2005–2007 гг., 2007–2009 гг., 2009–2011 гг., 2011–2013 гг.).

Каждый 60-секундный отрезок был представлен в виде двухмерной цифровой спектрограммы размером 1344×588 пиксель. Для получения спектрограмм использована находящаяся в открытом доступе программа для анализа и визуализации звуковых данных Sonic Visualiser 2.4.1.

Спектрограммы аудиозаписей Барака Обамы, сделанных в 2007 и 2014 гг., представлены на рис. 1. Вертикальное измерение спектрограммы соответствует частоте звукового отсчета в килогерцах, гори-



а *б*

Рис. 1. Спектрограмма аудиозаписей Барака Обамы, сделанных в 2007 (*а*) и 2014 г. (*б*)

зонтальное измерение — времени в секундах (0. . . 60 с). Следует отметить, что при анализе спектрограмм невооруженным глазом невозможно уловить различия между ними, хотя между записью первого аудиофрагмента и второго прошло 7 лет. Далее будет показано, что предложенный в настоящей статье метод способен проводить анализ спектрограмм и на его основе осуществлять хронологическое упорядочение аудиофрагментов.

Метод анализа данных. Анализ спектрограмм был проведен с использованием набора дескрипторов алгоритма Wndchrm, являющихся числовыми дескрипторами визуального контента (двухмерных спектрограмм) [12–15]. Предпосылкой для анализа является наблюдение, что визуальные свойства спектрограмм, например, границы и интенсивность пикселей, отображают аудиоданные в информативной манере [16, 17], а низкоуровневые свойства изображений спектрограмм могут быть эффективно использованы для классификациии отрезков речи и их упорядочения [18]. Алгоритм Wndchrm изначально разработан для проведения исследований в области биоинформатики [13] и признан эффективным при анализе двухмерных изображений в областях микроскопии и радиологии [19], астрономии [20], численном анализе предметов изобразительного искусства [21].

Алгоритм Wndchrm использует набор из 1025 двухмерных числовых дескрипторов визуальных данных. Они включают в себя, например, результаты преобразования Радона [22], фильтры Габора [23] с гауссовой гармонической функцией [24], многомасштабные гистограммы [25], оператор Прюитта [26] и др. Числовые дескрипторы извлекаются не только из непосредственно изображения спектрограммы, но также из его двухмерных преобразований и даже комбинаций преобразований более высоких порядков. Применяемые преобразо-

вания — преобразование Фурье, преобразование на основе полиномов Чебышева, вейвлет-преобразование, преобразование амплитуд границ.

Звук — сложный тип данных, поэтому эффективное численное представление звука часто требует большого числа параметров. Тем не менее, так как набор двухмерных числовых дескрипторов, полученных из каждой спектрограммы, является большим и полным, можно полагать, что не все из них одинаково информативны для анализа речевых отрезков.

Для оценки информативности дескрипторов каждому из них ставится в соответствие значение дискриминанта Фишера

$$W_f = \frac{\sum\limits_{c=1}^{N} \left(\overline{T_f} - \overline{T_{f,c}}\right)^2}{\sum\limits_{c=1}^{N} \sigma_{f,c}^2}, \qquad (1)$$

где N — число рассматриваемых временны́х интервалов; $\overline{T_f}$ — среднее значение числового дескриптора f во всем наборе входных данных; $\overline{T_{f,c}}$ и $\sigma_{f,c}^2$ — среднее значение и среднеквадратическое отклонение значения дескриптора f среди обучающего набора спектрограмм в пределах одного временно́го периода c. Все переменные в выражении (1) вычисляются после того, как значения числовых дескрипторов f нормализованы к интервалу [0, 1]. Когда каждому дескриптору поставлено в соответствие значение дискриминанта Фишера, 65 % дескрипторов с самыми малыми значениями дискриминанта Фишера отбрасываются. В результате получается набор из 154 числовых дескрипторов. В настоящей работе оптимальное значение порога 65 % было получено эмпирическим путем.

После вычисления вектора свойств дистанция $d_{x,c}$ между аудиофрагментом x и конкретным временны́м интервалом c рассчитывается по выражению

$$d_{x,c} = \frac{\sum\limits_{t \in T_c} \left[\sum\limits_{f=1}^{|x|} W_f (x_f - t_f)^2\right]^p}{|T_c|},$$

где T_c — обучающий набор для конкретного временно́го интервала c; t — вектор дескрипторов из набора T_c; $|x|$ — длина вектора дескриптора x; x_f — значение числового дескриптора f в векторе x_f; t_f — значение дескриптора f изображения t из обучающего набора; $|T_c|$ — число изображений в обучающем наборе периода c; p — показатель степени, $p = -5$ (это значение было подобрано эмпирическим путем). Дистанция между вектором дескрипторов конкретной спек-

трограммы в тестовом наборе и вектором дескрипторов конкретного временно́го интервала — среднее значение ее взвешенных дистанций до векторов дескрипторов всех отрезков речи, принадлежащих данному временно́му интервалу.

После вычисления дистанций между каждым отрезком речи из входного набора данных и всеми другими отрезками речи, дистанция $M_{A,Z}$ между временны́ми интервалами A и Z рассчитывается как средняя дистанция между всеми отрезками речи периода A и всеми отрезками речи периода Z попарно:

$$M_{A,Z} = \frac{\sum\limits_{s \in A} D_{s,Z}}{|A|},$$

где $|A|$ — число фрагментов речи, принадлежащих временно́му интервалу A.

Повторяя приведенные выше вычисления для всех периодов, в результате получаем матрицу дистанций между всеми периодами попарно. Таким образом, в ячейке $[n, m]$ матрицы содержится значение дистанции между временны́ми интервалами n и m. Из матрицы дистанций получаем матрицу сходства, при этом элементы матрицы нормализуются так, что вычисленная дистанция от конкретного периода до каждого другого периода делится на вычисленную дистанцию от этого периода до самого себя (следовательно, значение сходства периода с самим собой устанавливается в единице).

Во всех экспериментах, описаннных далее, несколько речевых отрезков каждого временно́го интервала было использовано для тестирования метода, а остальные — для обучения. Каждый эксперимент был повторен 40 раз, при этом во время каждого прохода метода аудиоотрезки были произвольно распределены между обучающим и тестовым наборами.

Необходимо отметить недостаток метода — время работы. Извлечение вектора двухмерных числовых дескрипторов из одной спектрограммы занимает порядка 6 мин (в ходе эксперимента был использован компьютер с процессором Intel Core i7).

Результаты экспериментов. В ходе первого эксперимента были проанализированы аудиофайлы — выступления Барака Обамы. Данные для эксперимента (аудиофайлы) были взяты с сайта www.americanrhetoric.com. Для каждого периода было взято 14 аудиофрагментов, 11 из которых были использованы для обучения, а оставшиеся три — для тестирования. Аудиоотрезок представлял собой отрывок длительностью 1 мин из выступления Обамы. Эксперимент был повторен 40 раз с произвольным размещением входных аудиофайлов в обучающий или тестовый наборы. Точность определения принадежности речевого отрезка к требуемому временно́му интервалу составила 59 %. Матрица

сходства классов, вычисленная в результате работы метода, представлена в табл. 1.

Таблица 1

Матрица сходства (эксперимент 1)

Период	Период			
	1	2	3	4
1	1,00	1,00	0,93	0,93
2	0,97	1,00	0,98	0,96
3	0,79	0,80	1,00	0,96
4	0,67	0,69	0,78	1,00

Дерево сходства (рис. 2, *а*) для периодов, рассматриваемых в эксперименте 1, было построено с помощью программы “Графоанализатор” (www.grafoanalizator.unick-soft.ru). Отправной точкой является класс “1 период”. Оставшиеся классы расположены так, чтобы расстояние до класса было обратно пропорционально значению сходства класса относительно класса “1 период”. Таким образом, чем больше расстояние до класса, тем меньше значение сходства между данным классом и классом “1 период”, т.е. метод определил указанные классы как менее схожие друг с другом.

Метод смог расположить временны́е интервалы в хронологическом порядке. Так, согласно данным, приведенным в табл. 1, наиболее схожим с классом “1 период” является класс “2 период” (значение сходства равно 0,985), затем в порядке убывания значения сходства следуют класс “3 период” (значение сходства равно 0,86) и класс “4 период” (значение сходства равно 0,8).

Числовые дескрипторы с самым высоким значением дискриминанта Фишера для этого эксперимента (дескрипторы, которые оказались самыми существенными в процессе работы метода) следующие:

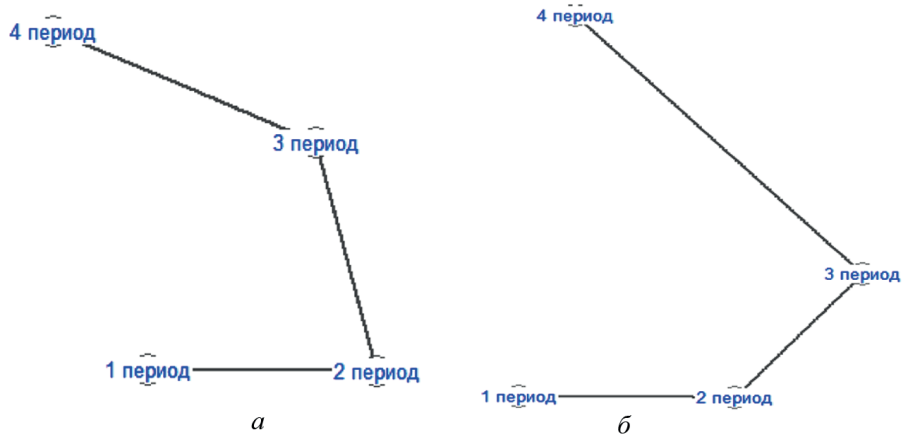


Рис. 2. Дерево сходства для экспериментов 1 (*а*) и 2 (*б*)

— Comb4Orient4MomentsHistogram_Wavelet 0_Kurt_HistBin00: 4.716667;
— Comb4Orient4MomentsHistogram_Wavelet 0_Kurt_HistBin00: 4.716667;
— Comb4Orient4MomentsHistogram_ChebyshevFFT Plus45_Skew_HistBin00;
— 4.259768;
— MultipleScaleHistograms_TBins3_Bin01: 3.505789;
— MultipleScaleHistograms_TBins5_Bin02: 3.061456.

В эксперименте 2 были проанализированы аудиофайлы — выступления канцлера Германии Ангелы Меркель, при этом набор входных данных был разбит на четыре временны́х интервала:

— 22 ноября 2005 г.–22 ноября 2007 г.;
— 22 ноября 2007 г.–22 ноября 2009 г.;
— 22 ноября 2009 г.–22 ноября 2011 г.;
— 22 ноября 2011 г.–22 ноября 2013 г.

Данные для эксперимента (аудиофайлы) были взяты с официального сайта Бундестага (www.bundestag.de). Для каждого периода выбрано 13 аудиофрагментов, 10 из которых были использованы для обучения, а оставшиеся три — для тестирования. Аудиоотрезок представлял собой отрывок длительностью 1 мин из выступления Меркель. Эксперимент был повторен 40 раз с произвольным размещением входных аудиофайлов в обучающий или тестовый наборы. Метод точно определил принадлежность аудиофрагмента к требуемому периоду в 68 % случаев. Указанное значение выше, чем полученное значение для любого ранее проведенного эксперимента. Это свидетельствует о том, что в случае с Меркель характеристики речи имеют наиболее заметную динамику на протяжении всего рассматриваемого периода (8 лет). Матрица сходства классов, вычисленная в результате работы метода, представлена в табл. 2.

Таблица 2

Матрица сходства (эксперимент 2)

Период	Период			
	1	2	3	4
1	1,00	1,02	0,91	0,69
2	0,86	1,00	0,89	0,50
3	0,77	0,91	1,00	0,58
4	0,48	0,54	0,54	1,00

Точность определения принадлежности речевого отрезка к требуемому временно́му интервалу может быть повышена, если входные

данные будут подвергнуты предварительной обработке (включает в себя предкоррекцию или выравнивание спектра сигнала, фильтрацию шума, логарифмическое сжатие спектра, нормализацию звука).

Дерево сходства для периодов, рассматриваемых в эксперименте 2, показано на рис. 2, *б*. Метод расположил временны́е интервалы в правильном хронологическом порядке. Согласно данным, приведенным в табл. 2, наиболее схожим с классом “1 период” является класс “2 период” (значение сходства равно 0,94), затем в порядке убывания значения сходства следуют класс “3 период” (значение сходства равно 0,84) и класс “4 период” (значение сходства равно 0,585).

Числовые дескрипторы с самым высоким значением дискриминанта Фишера для этого эксперимента следующие:

— MultipleScaleHistograms_TBins3_Bin01: 11.098907;
— ChebyshevCoefficientHistogram_Bin20: 9.000000;
— ZernikeMoments_Z_03_03: 8.974037;
— ZernikeMoments_Z_03_01: 8.200648;
— MultipleScaleHistograms_TBins5_Bin02: 7.810644.

Заключение и выводы. Как было отмечено выше, звук является сложным типом данных, если рассматривать его с позиции автоматического анализа с помощью вычислительных машин. В настоящей статье был описан метод, который использует автоматический анализ спектрограмм аудиофрагментов для построения матрицы сходства между разными временны́ми интервалами.

Результаты экспериментов показали, что предложенный метод способен расположить временны́е интервалы в хронологическом порядке, т.е. он способен отслеживать изменения в характеристиках речи человека, произошедшие за определенный период времени (в эксперименте рассмотрены периоды длительностью 8–14 лет). Также была исследована чувствительность метода для данных, разделенных на большее число коротких периодов. Результаты показали, что, несмотря на снижение точности определения принадлежности речевого отрезка к требуемому временно́му интервалу, метод смог расположить временны́е периоды в хронологическом порядке. Полученные результаты показывают, что автоматический анализ спектрограмм может быть эффективно использован для анализа звука.

Точность определения принадлежности речевого отрезка к требуемому временно́му интервалу может быть повышена, если входные данные будут подвергнуты предварительной обработке. Точность работы метода можно повысить варьированием размера вектора свойств, используемого для анализа, и продолжительности выбранных аудиофрагментов. При этом следует учитывать, что с увеличением размера вектора линейно возрастает и время работы метода.

В практических компьютерных приложениях рассмотренный метод может быть полезен для организации и упорядочения аудиоданных (например, для автоматического создания аудиоархивов), для анализа и визуализации сходства в характеристиках речи при проведении различных исследований (идентификация человека по голосу, обнаружение похожих голосов и т.д.).

ЛИТЕРАТУРА

1. *Tzanetakis G.*, *Cook P.* Musical genre classification of audio signals // IEEE Transactions on Speech and Audio Processing. 2002. Vol. 10. P. 293–302.
2. *Guo G.*, *Li S.Z.* Content-based audio classification and retrieval by support vector machines // IEEE Transactions on Neural Networks. 2003. Vol. 14. P. 209–215.
3. *Li T.*, *Ogihara M.*, *Li Q.* A comparative study on content-based music genre classification // SIGIR03. 2003. P. 282–289.
4. *Bagci U.*, *Erzin E.* Automatic Classification of Musical Genres Using Inter-Genre Similarity // IEEE Signal Processing Letters. 2007. Vol. 14. P. 521–524.
5. *Toward* multi-modal music emotion classification / Y.H. Yang et al. // Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing. 2008. P. 70–79.
6. *Zlatintsi A.*, *Maragos P.* Multiscale fractal analysis of musical instrument signals with application to recognition // IEEE Transactions on Audio, Speech and Language Processing. 2013. Vol. 21. P. 737–748.
7. *McFee B.*, *Barrington L.*, *Lanckriet G.R.G.* Learning content similarity for music recommendation// IEEE Transactions on Audio, Speech and Language Processing. 2012. Vol. 20. P. 2207–2218.
8. *Predictability* of music descriptor time series and its application to cover song detection / Y. Serra et al. // IEEE Transactions on Audio, Speech and Language Processing. 2012. Vol. 20. P. 514–525.
9. *Manders A.J.*, *Simpson D.M.*, *Bell S.L.* Objective prediction of the sound quality of music processed by an adaptive feedback canceller // IEEE Transactions on Audio, Speech and Language Processing. 2012. Vol. 20. P. 1734–1745.
10. *Downie D.* The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research // Acoustical Science and Technology. 2008. Vol. 29. P. 247–255.
11. *Casey M.* et al. Content-based music information retrieval: Current directions and future challenges // Proceedings of the IEEE. 2008. Vol. 96. P. 668–695.
12. *George J.*, *Shamir L.* Computer analysis of similarities between albums in popular music // Pattern Recognition Letters. 2014. Vol. 45. P. 78–84.
13. *Wndchrm* — an open source utility for biological image analysis / L. Shamir et al. // Source Code For Biology And Medicine. 2008. URL: http://www.scfbm.org/content/3/1/13 (дата обращения: 01.10.2014).
14. *Shamir L.* Evaluation of face datasets as tools for assessing the performance of face recognition methods // International Journal of Computer Vision. 2008. Vol. 79. P. 225–230.
15. *WND-CHARM*: Multipurpose image classification using compound image transforms / N. Orlov et al. // Pattern Recognition Letters. 2008. Vol. 29. P. 1684–1693.
16. *Deshpande H.*, *Singh R.*, *Nam U.* Classification of music signals in the visual domain // Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01). 2001. Vol. 1. P. 1–10.

17. *Holzapfel A.*, *Stylianou Y.* Musical genre classification using nonnegative matrix factorization-based features // IEEE Transactions on Audio, Speech and Language Processing. 2008. Vol. 16. P. 424–434.
18. *Music* genre recognition using spectrograms / Y.M.G. Costa et al. // 18th International Conference on Systems, Signals and Image Processing. 2011. P. 1–4.
19. *IICBU 2008* — A proposed benchmark suite for biological image analysis / L. Shamir et al. // Source Code for Biology and Medicine. 2008. Vol. 46. P. 943–947.
20. *Shamir L.* Automatic morphological classification of galaxy images // Monthly Notices of the Royal Astronomical Society. 2009. Vol. 399. P. 1367–1372.
21. *Shamir L.* Computer analysis reveals similarities between the artistic styles of Van Gogh and Pollock // Leonardo. 2012. Vol. 45. P. 149–154.
22. *Lim J.S.* Two-Dimensional signal and image processing // Prentice Hall. 1990. P. 42–45.
23. *Gabor D.* Theory of communication // Journal of IEEE. 1946. Vol. 93. P. 429–457.
24. *Gregorescu C.*, *Petkov N.*, *Kruizinga P.* Comparison of texture features based on Gabor filters // IEEE Transactions on Image Processing. 2002. Vol. 11. P. 1160–1167.
25. *Hadjidementriou E.*, *Grossberg M.*, *Nayar S.* Spatial information in multiresolution histograms// IEEE Conference on Computer Vision and Pattern Recognition. 2001. Vol. 1. P. 702.
26. *Prewitt J.M.* Object enhancement and extraction. Picture processing and psychopictoris // Academic Press. 1970. P. 75–149.

REFERENCES

[1] Tzanetakis G., Cook P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 2002, vol. 10, pp. 293–302.
[2] Guo G., Li S.Z. Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, 2003, vol. 14, pp. 209–215.
[3] Li T., Ogihara M., Li Q. A comparative study on content-based music genre classification. *SIGIR 03*, 2003, pp. 282–289.
[4] Bagci U., Erzin E. Automatic Classification of Musical Genres Using Inter-Genre Similarity. *IEEE Signal Processing Letters*, 2007, vol. 14, pp. 521–524.
[5] Yang Y.H. et al. Toward multi-modal music emotion classification. *Proceedings of the 9th Pacific Rim Conference on Multimedia*: *Advances in Multimedia Information Processing*, 2008, pp. 70–79.
[6] Zlatintsi A., Maragos P. Multiscale fractal analysis of musical instrument signals with application to recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2013, vol. 21, pp. 737–748.
[7] McFee B., Barrington L., Lanckriet G.R.G. Learning content similarity for music recommendation. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, vol. 20, pp. 2207–2218.
[8] Serra Y. et al. Predictability of music descriptor time series and its application to cover song detection. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, vol. 20, pp. 514–525.
[9] Manders A.J., Simpson D.M., Bell S.L. Objective prediction of the sound quality of music processed by an adaptive feedback canceller. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, vol. 20, pp. 1734–1745.
[10] Downie D. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 2008, vol. 29, pp. 247–255.
[11] Casey M. et al. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 2008, vol. 96, pp. 668–695.

[12] George J., Shamir L. Computer analysis of similarities between albums in popular music. *Pattern Recognition Letters*, 2014, vol. 45, pp. 78–84.

[13] Wndchrm — an open source utility for biological image analysis / L. Shamir et al. // Source Code For Biology And Medicine. 2008. URL: http://www.scfbm.org/content/3/1/13 (accessed: 01.10.2014).

[14] Shamir L. Evaluation of face datasets as tools for assessing the performance of face recognition methods. *International Journal of Computer Vision*, 2008, vol. 79, pp. 225–230.

[15] Orlov N. et al. WND-CHARM: Multipurpose image classification using compound image transforms. *Pattern Recognition Letters*, 2008, vol. 29, pp. 1684–1693.

[16] Deshpande H., Singh R., Nam U. Classification of music signals in the visual domain. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01)*, 2001, vol. 1, pp. 1–10.

[17] Holzapfel A., Stylianou Y. Musical genre classification using nonnegative matrix factorization-based features. *IEEE Transactions on Audio, Speech and Language Processing*, 2008, vol. 16, pp. 424–434.

[18] Costa Y.M.G. et al. Music genre recognition using spectrograms. *18th International Conference on Systems, Signals and Image Processing*, 2011, pp. 1–4.

[19] Shamir L. et al. IICBU 2008 — A proposed benchmark suite for biological image analysis. *Source Code for Biology and Medicine*, 2008, vol. 46, pp. 943–947.

[20] Shamir L. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 2009, vol. 399, pp. 1367–1372.

[21] Shamir L. Computer analysis reveals similarities between the artistic styles of Van Gogh and Pollock. *Leonardo*, 2012, vol. 45, pp. 149–154.

[22] Lim J.S. Two-Dimensional signal and image processing. *Prentice Hall*, 1990, pp. 42–45.

[23] Gabor D. Theory of communication. *Journal of IEEE*, 1946, vol. 93, pp. 429–457.

[24] Gregorescu C., Petkov N., Kruizinga P. Comparison of texture features based on Gabor filters. *IEEE Transactions on Image Processing*, 2002, vol. 11, pp. 1160–1167.

[25] Hadjidementriou E., Grossberg M., Nayar S. Spatial information in multiresolution histograms. *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, p. 702.

[26] Prewitt J.M. Object enhancement and extraction.Picture processing and psychopictoris. *Academic Press*, 1970, pp. 75–149.

Статья поступила в редакцию 27.01.2015

Алфимцев Александр Николаевич — канд. техн. наук, доцент кафедры "Информационные системы и телекоммуникации" МГТУ им. Н.Э. Баумана. Автор более 80 научных работ, в том числе пяти охранных документов на интеллектуальную собственность, в области методов искусственного интеллекта, мультимодальных интерфейсов и распознавания образов.
МГТУ им. Н.Э. Баумана, Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5.

Alfimtsev A.N. — Cand. Sci. (Eng.), assoc. professor of the Information Systems and Telecommunications Department of the Bauman Moscow State Technical University. Author of more than 80 publications including five patents for inventions in the fields of artificial intelligence methods, multimodal interfaces and pattern recognition.
Bauman Moscow State Technical University, 2-ya Baumanskaya ul. 5, Moscow, 105005 Russian Federation.

Назарова Светлана Игоревна — инженер компании ОАО “Газпром автоматизация”. Автор трех научных работ в области мультимодальных интерфейсов и распознавания образов.
ОАО “Газпром автоматизация”, Российская Федерация, 119435, Москва, Саввинская наб., д. 25.

Nazarova S.I. — engineer at OAO “Gazprom Automation”. Author of three publications in the fields of multimodal interfaces and pattern recognition.
OAO “Gazprom Automation”, Savvinskaya nab. 25, Moscow, 119435 Russian Federation.

Просьба ссылаться на эту статью следующим образом:

Алфимцев А.Н., Назарова С.И. Хронологическое упорядочение аудиофрагментов с использованием двухмерных спектрограмм // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. 2015. № 3. С. 127–139.

Please cite this article in English as:

Alfimtsev A.N., Nazarova S.I. Chronological ordering of the audio data using 2D spectrograms. *Vestn. Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Priborostr.* [Herald of the Bauman Moscow State Tech. Univ., Instrum. Eng.], 2015, no. 3, pp. 127–139.

Вниманию авторов журнала “Вестник МГТУ им. Н.Э. Баумана. Серия “Приборостроение”

Редакция журнала принимает к рассмотрению статьи, оформленные в соответствии с действующими правилами, по следующей тематике.

Приборостроение, метрология и информационно-измерительные приборы и системы

- Приборы и методы измерения
- Приборы навигации
- Акустические приборы и системы
- Оптические и оптико-электронные приборы и комплексы
- Радиоизмерительные приборы
- Приборы и методы для измерения ионизирующих излучений и рентгеновские приборы
- Приборы и методы контроля природной среды, веществ, материалов и изделий
- Технология приборостроения
- Метрология и метрологическое обеспечение
- Информационно-измерительные и управляющие системы
- Приборы, системы и изделия медицинского назначения
- Приборы и методы преобразования изображений и звука

Радиотехника и связь

- Радиотехника, в том числе системы и устройства телевидения
- Антенны, СВЧ-устройства и их технологии
- Системы, сети и устройства телекоммуникаций