ИНФОРМАТИКА И УПРАВЛЕНИЕ В ТЕХНИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМАХ

УДК 004.931

В. Е. Гай, И. А. Пресняков, М. И. Арабаджи,
М. О. Дербасов, И. В. Поляков, Е. Н. Викулова

АНАЛИЗ АУДИОЗАПИСЕЙ С ПОЗИЦИЙ ТЕОРИИ АКТИВНОГО ВОСПРИЯТИЯ

Нижегородский государственный технический университет им. Р. Е. Алексеева

Рассматриваются проблемы автоматизации процессов анализа звуковых записей. Информация, выделенная из аудиозаписи, может содержать полезные сведения, в данном случае приводятся примеры создания системы определения пола диктора по голосу, а также системы нахождения схожих по эмоциональному составу аудиозаписей. Целью работы является создание систем анализа записей звука, используя распознавание образов. На одном из этапов распознавания предлагается применить теорию активного восприятия, что позволит добиться высокой точности работы при уменьшении вычислительных затрат.

Ключевые слова: распознавание образов, машинное обучение, теория активного восприятия, аудиозаписи, звуковые сигналы, идентификация диктора, система рекомендаций.

Введение

С ростом автоматизации производственных процессов возникает необходимость минимизировать участие человека при работе с ними. Главным образом это касается тех сфер деятельности, где работа однообразна либо опасна для человека. В таких процессах человека целесообразно заменить на автоматические системы, способные реагировать на различные отклонения параметров технологического процесса от номинальных. Пройдя развитие от распознающих автоматов до современных интеллектуальных систем распознавания, точность работы современных устройств позволяет использовать их в большом количестве сфер деятельности. При замене человека такой системой возможно повышение качества результатов (при работе с многократно повторяющимися рутинными операциями), а также скорости выполнения различных операций, что особенно критично при обстоятельствах, влияющих на психологическое состояние человека, принимающего решение.

При построении современных систем, решающих задачи распознавания образов, часто используются методы машинного обучения. Данные методы применяются на практике при решении задач в сферах медицинской диагностики, экономике, робототехнике, при распознавании речи, текста, рукописного ввода, а также в компьютерном зрении. Конкретное применение в области анализа музыкальных композиций и записей человеческой речи представлено далее.

При решении задач распознавания выделяют три этапа: предварительная обработка (формирование исходного описания); формирование системы признаков; классификация (принятие решения). На первом этапе к исходному сигналу применяются различные фильтры для избавления от шумов, также сигнал разделяется на сегменты определенной длительности. На этапе формирования системы признаков используются такие методы, как мел-

частотные кепстральные коэффициенты, вейвлет-коэффициенты, Коэффициенты оконного преобразования Фурье, LPC-коэффициенты, скрытая марковская модель, модель гауссовой смеси, методы глубокого обучения и др. Наконец, на этапе классификации используются байесовские классификаторы (методы опорных векторов, метод ближайших соседей), линейные разделители, нейронные сети и др.

1. Распознавание образов с позиций теории активного восприятия

2.1. Введение в теорию активного восприятия

Систему анализа аудиозаписи можно представить как систему распознавания образов. Известно, что данная система образов включает три этапа обработки данных: предварительная обработка, вычисление признаков и принятие решения [2]. Для реализации этапов предварительной обработки и вычисления признаков звукового сигнала предлагается использовать теорию активного восприятия [3].

Предварительная обработка заключается в выполнении Q-преобразования, т.е. в применении к сегментам исходного сигнала операции сложения:

$$g(t) = \sum_{k=(t-1)\cdot L+1}^{t\cdot L} f(k), t = \overline{1, N},$$

где L – число отсчётов, входящих в сегмент; N – число сегментов сигнала; g – результат применения Q-преобразования к сигналу f; $f(k)$ – k-й отсчёт сигнала f; $g(t)$– t-й отсчёт сигнала g.

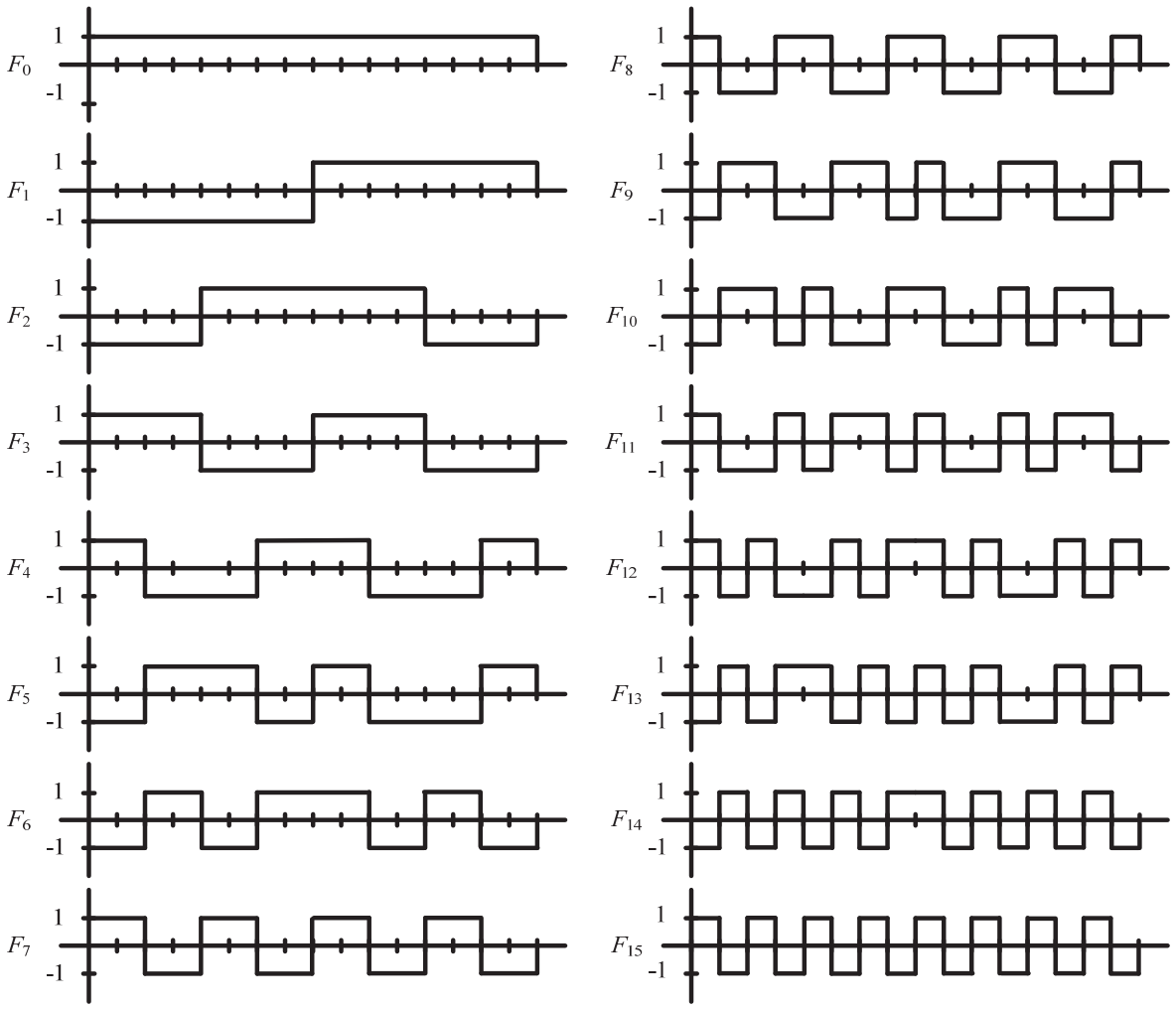

Рис. 1. Используемые фильтры

Формирование признакового описания исходного сигнала заключается в применении к сигналу g множества фильтров Уолша системы Хармута (рис. 1, показано 16 фильтров):

$$\mu(k, c(t)) = \sum_{i=0}^{M-1} F_k(i) \cdot g(((t-1) \cdot M + 1) : (t \cdot M)),$$

где $\mu(k, c(t))$ – результат применения множества фильтров Уолша системы Хармута к сигналу g; $k = \overline{0, M-1}$; $t = \overline{0, |c|-1}$; $c = \{1, P, 2 \cdot P, 3 \cdot P, ..., N - T \cdot P\}$ – множество значений смещений по сигналу g, $|c|$ – мощность множества c, P – величина смещения по сигналу g ($1 \le P \le M$), M – число используемых фильтров. Таким образом, признаковое описание сигнала представляет собой матрицу размером $M \times |c|$, причём каждая строка признакового описания представляет собой результат U-преобразования сегмента сигнала.

Последовательное применение к сигналу Q-преобразования и системы фильтров реализуют U-преобразование, являющееся базовым в теории активного восприятия. U-преобразование имеет минимально возможную вычислительную сложность, поскольку при его реализации используются простейшие операции – сложение и вычитание. Стандартные преобразования требуют реализации свертки, а на уровне весовых коэффициентов – операции арифметического умножения.

Теория активного восприятия не ограничивается только формированием спектрального представления сигнала. В состав теории входит раздел «Алгебра групп», посвящённый анализу зависимостей между спектральными коэффициентами разложения. Обнаруженные зависимости допускают своё использование на этапах принятия решения и понимания анализируемого сигнала. Пусть каждому фильтру $F_i \in \{F_i\} \equiv \boldsymbol{F}$ соответствует координатно-определенный бинарный оператор $V_i \in \{V_i\} \equiv \boldsymbol{V}$; тогда компоненте $\mu_i \neq 0$ вектора μ допустимо поставить в соответствие оператор V_i либо $\overline{V_i}$ (в зависимости от знака компоненты). В результате вектору μ ставится в соответствие подмножество операторов из $\{V_i\}$, имеющих аналогичную фильтрам конструкцию, но разное значение элементов матрицы ($+1 \leftrightarrow 1$; $-1 \leftrightarrow 0$). Задавая на множестве $\{V_i\}$ операции теоретико-множественного умножения и сложения, имеем алгебру описания изображения в двумерных булевых функциях. С учётом инверсий всего существует 15 операторов, которые могут использоваться при формировании признакового описания, так как оператор V_0 принимает только прямое значение.

На множестве операторов формируется алгебра групп (этап синтеза) анализируемого сигнала:

- семейство алгебраических структур (названных полными группами) $\{P_{ni}\}$ вида $P_{ni} = \{V_i, V_j, V_k\}$ мощности 35;
- семейство алгебраических структур (названных замкнутыми группами) $\{P_{si}\}$ вида $P_{si} = \{V_i, V_j, V_k, V_r\}$ мощности 105, где каждая группа образована из пары определенным образом связанных полных групп.

Среди полных групп выделяют полные группы на операции сложения и на операции умножения, среди замкнутых групп – замкнутые группы и замкнутые множества.

Две группы (полные или замкнутые) называются несовместными, если в их состав входят операторы с одинаковыми номерами, но с разными знаками.

С помощью замкнутых и полных групп выполняется спектрально-корреляционный анализ. Полные группы позволяют выявить корреляционные связи между операторами, замкнутые – выявить корреляционные связи между полными группами. Если множество операторов – алфавит, то множества групп – более сложные грамматические описания наблюдаемого сигнала: полная группа – слово, замкнутая группа – словосочетание.

Используя спектральное представление сигнала μ, формируется множество операторов, описывающих данный сигнал, а затем множества полных и замкнутых групп:

$V = GV[\mu]$, $P_{na} = GP_{na}[\mu, V]$, $P_{nm} = GP_{nm}[\mu, V]$,

$P_s = GP_s[\mu, V, P_{na}, P_{nm}]$, $P_c = GP_c[\mu, V, P_{na}, P_{nm}]$,

где GV – оператор вычисления по спектральному представлению сигнала признакового описания V на основе операторов; GP_{na} (GP_{nm}) – на основе полных групп на операции сложения P_{na} (умножения, P_{nm}); $GP_c(GP_s)$ – на основе замкнутых групп P_s (замкнутых множеств, P_c).

1.2. Реализация этапов системы распознавания

Используя признаковые описания, можно получить интегрированные признаковые описания в виде гистограмм частоты появления операторов, полных и замкнутых групп (двумерные, трёхмерные):

$$h_V = H[V, \Gamma], h_{na} = H[P_{na}, \Gamma], h_{nm} = H[P_{nm}, \Gamma], h_s = H[P_s, \Gamma], h_c = H[P_c, \Gamma],$$

где h_{V} – гистограмма операторов; h_{na} (h_{nm}) – гистограмма полных групп на операции сложения (умножения); $h_{\mathrm{s}}(h_c)$ – гистограмма замкнутых множеств (замкнутых групп); Γ – размерность гистограммы: $1d$ – одномерная гистограмма; $2d$ – двумерная гистограмма; $3d$ – трёхмерная гистограмма. В двумерной гистограмме учитываются возможные появления пар групп в описании одного сегмента сигнала, в трёхмерной – троек.

В табл. 1 приведена оценка размерности предлагаемых интегрированных признаковых описаний (размерности приведены с учётом возможных инверсий операторов, включённых в полные и замкнутые группы).

Таблица 1

Размерность систем признаков

Система признаков / Размерность	$1d$	$2d$	$3d$
h_V	1 × 30	30 × 30	30 × 30 × 30
$h_{na}(h_{nm})$	1 × 140	140 × 140	140 × 140 × 140
$h_s(h_c)$	1 × 840	840 × 840	840 × 840 × 840

При увеличении размерности предложенных систем признаков количество признаков растёт как показательная функция. В связи с этим было принято решение ограничиться только двумерными гистограммами групп и трёхмерной гистограммой операторов. Двумерные гистограммы, представляющие собой матрицы, обладают следующими свойствами:

- нулевой главной диагональю;
- симметрией относительно главной диагонали.

Таким образом, число незначащих элементов в двумерной гистограмме составляет $(H_{dim} \cdot H_{dim} - H_{dim}) / 2$, где H_{dim} – число столбцов (строк) матрицы.

Таблица 2

Оценка размерности систем признаков до и после сжатия

Система признаков / Размерность	До сжатия			После сжатия		
	$1d$	$2d$	$3d$	$1d$	$2d$	$3d$
h_V	1 × 30	30 × 30	30 × 30 × 30	-	1 × 435 (2,06)	-
$h_{na}(h_{nm})$	1 × 140	140 × 140	140 × 140 × 140	-	1 × 5050 (3,88)	-
$h_s(h_c)$	1 × 840	840 × 840	840 × 840 × 840	-	1 × 106030 (6,65)	-

Учитывая несовместность некоторых полных и замкнутых групп возможно сокращение размерности предложенных систем признаков. Сжатие с учётом несовместности возможно только для гистограмм размерностью больше или равных двум. Размерности систем признаков на основе гистограмм показаны в табл. 2. В скобках указан коэффициент сжатия, рассчитываемый как отношение числа признаков до сжатия к числу признаков после сжатия.

Схема формирования интегрированных признаковых описаний показана на рис. 2.

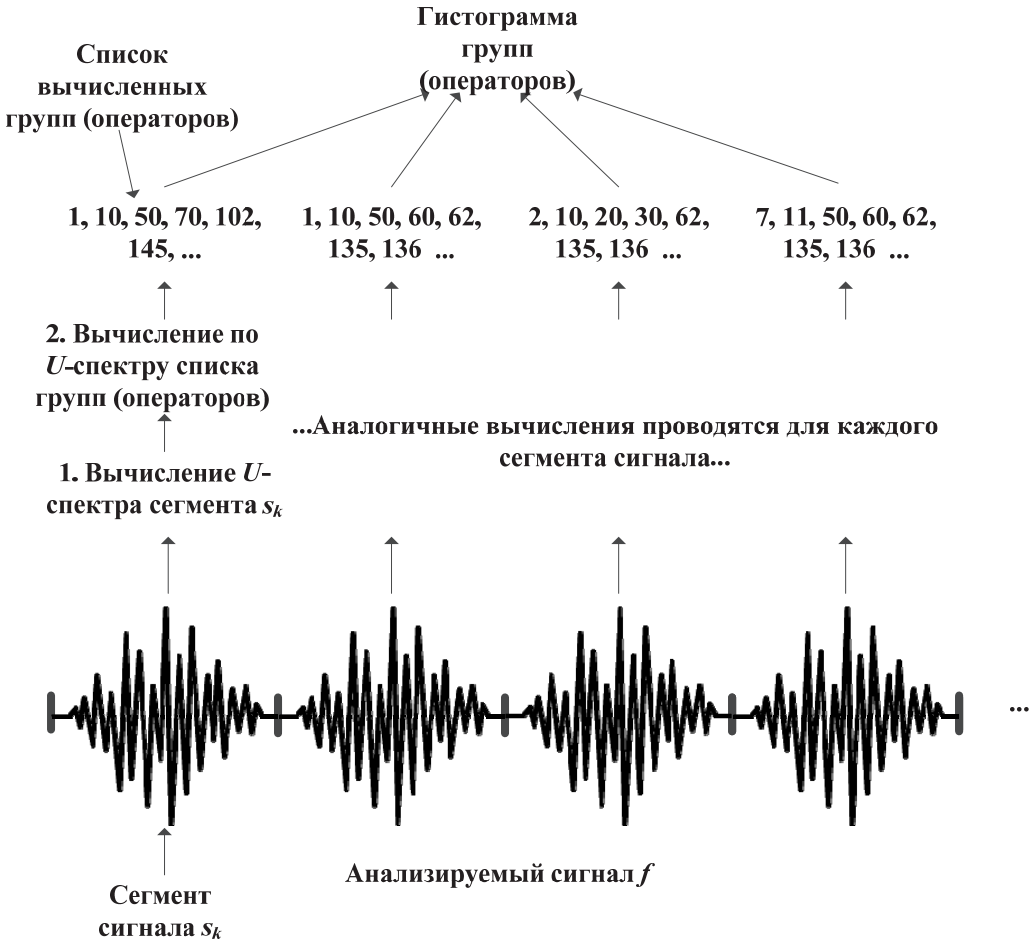

Рис. 2. Схема формирования системы признаков

2. Реализация рекомендательной системы подбора аудиозаписей

С развитием глобальных сетей большое распространение получили социальные сети. Такие сайты содержат множество мультимедийной информации (музыка, видео, изображения). Но так как поток информации слишком большой, пользователям обычно сложно найти что-то интересное именно для них. Чтобы помочь людям с этой проблемой и были созданы рекомендательные системы. Их задача заключается в прогнозировании того, какой контент будет интересен человеку, имея информацию о его профиле. Множество социальных сетей уже имеют рекомендательные системы, например, Яндекс музыка, AppleMusic, Вконтакте. Все они работают по схожему принципу. В них применяется метод коллаборативной фильтрации - в этом методе оценивается поведение пользователей в прошлом: их оценки той или иной музыке, покупки и частота прослушиваний тех или иных треков. Система сравнивает профили всех пользователей. Это делается для выявления людей со схожими музыкальными предпочтениями: что нравится одному, может понравиться и другому [4]. Однако системы такого рода не лишены недостатков, основным из которых является «холодный старт», т.е. когда рекомендательная система только начинает свою работы и пользователи еще не успели совершить достаточно действий, результат работы системы будет посредственным, так как ей не на чем основывать свои рекомендации.

Описанная далее система построена на принципе фильтрации содержимого – признаки для рекомендации выстраиваются не на основе действий пользователей, а на признаках, содержащихся в контенте. Это позволяет избежать проблемы «холодного старта», так как нет необходимости ожидать действий пользователей для генерации рекомендаций. В качест-

ве признака в такой системе выступает эмоциональный отклик. То есть в качестве признаков классификации в данной системе выступают эмоции, которые музыка вызывает у людей. Такая система позволяет иначе оценивать контент и давать рекомендации, которые классические системы не могу дать.

Так как центральной функцией системы является получение эмоционального состава аудиозаписей, в данном случае она рассматривается с точки зрения системного подхода. Задача состоит из этапов предварительной обработки сигналов, вычисления признаков и классификации. Для этапов предварительной обработки сигнала и формирования системы признаков было решено использовать теорию активного восприятия, а для этапа классификации – метод опорных векторов. Метод опорных векторов – алгоритм для классификации, основанный на обучении с учителем. Идея метода состоит в переводе исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве [5].

Классификатор SVM необходимо обучить перед использованием, таким образом, было принято решение сделать его обучение в виде отдельной программы и сохранять созданный ей классификатор.

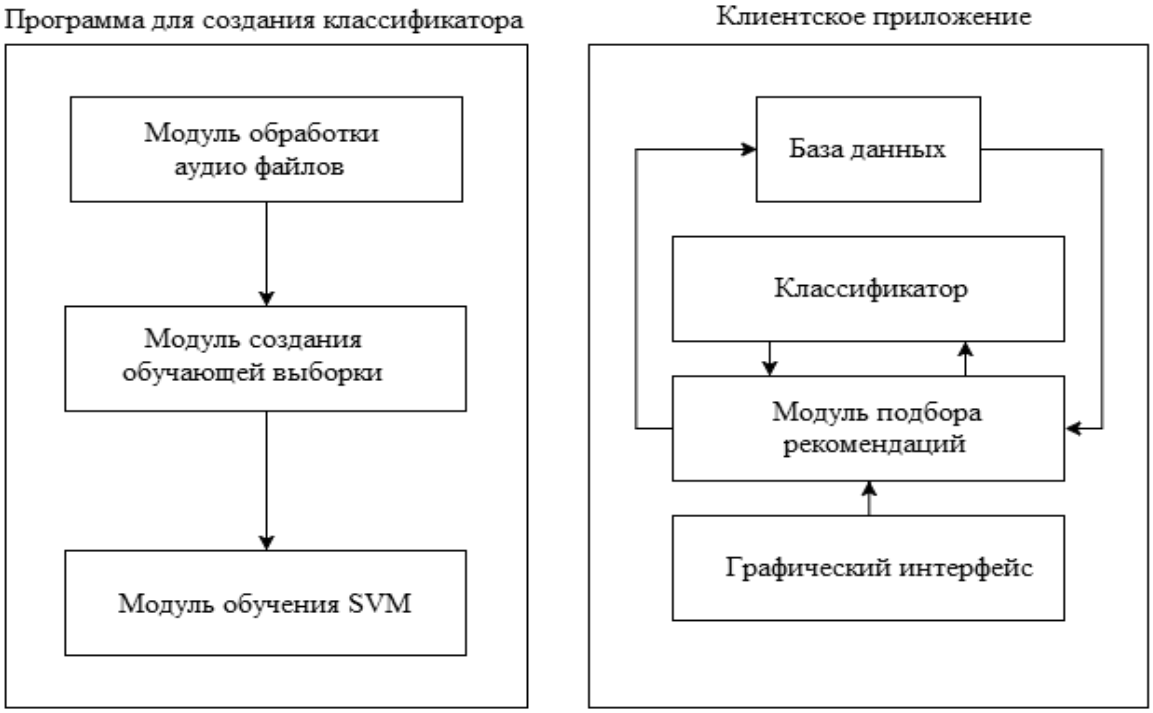

Рис. 3. Структура приложения

В качестве входных данных программа использует заранее подготовленный набор файлов, разбитых по пяти разным эмоциям, которые они вызывают. Далее программа открывает каждый файл, делит его на 10 частей, продолжительностью по 5 с и над каждой из частей производит операции *U*-преобразования, вычисления замкнутых групп на основе *U*-преобразования и добавления полученных признаков в общий массив. Далее необходимо обучить используемый классификатор с помощью полученной выборки.

Основная задача клиентского приложения – получить от пользователя аудиозаписи, вычислить их эмоциональный состав и найти в базе данных похожие записи. Таким образом, его алгоритм работы следующий:

1. Получение признаков из файла, полученного из графического интерфейса – файл подвергается такой же обработке, как и при создании классификатора (файл делится на 10 частей, выполняется *U*-преобразование, высчитываются закрытые группы);

2. Подключение классификатора – программа загружает заранее подготовленный в первом приложении классификатор для использования;

3. Признаки, полученные на первом этапе, передаются в классификатор. В итоге получаем эмоциональный состав этой аудиозаписи;

4. В базе данных происходит поиск наиболее похожих по эмоциональному составу записей с той, которую передал пользователь. Получаем список записей, рекомендованных пользователю, которые передаем в графический интерфейс;

5. Состав текущей записи добавляется в базу данных, чтобы она могла быть рекомендована в дальнейшем.

Для экспериментальной оценки эффективности системы был использован следующий метод: была создана тестовая подборка аудиозаписей, для которых были заранее определены классы. В данном случае подборка состояла из 25 музыкальных файлов – по пять файлов на каждый класс.Эти файлы были переданы в обученный классификатор. Далее в полученном эмоциональном составе для каждой записи выбиралась доминирующая эмоция, она сравнивалась с заранее заданным классом – если они равны, то система определила класс верно. Для тестовой выборки классы 19 из 25 файлы были определены верно.

3. Реализация идентификации пола диктора по голосу

Задача идентификации пола диктора по голосу – одна из подзадач актуальной в последнее время проблемы идентификации личности человека по физиологическим особенностям его голоса. Решение этой проблемы применимо в различных системах, связанных с распознаванием речи. В задачу идентификации личности входит оценка возраста диктора, его эмоционального состояния, среднего тона голоса и т. д. Применение задачи идентификации личности по голосу в системах, реализующих перевод речи в текст или распознавание голосовых команд, необходимо для устранения ошибок распознавания, причиной которых являются не учтённые заранее физиологические особенности голоса говорящего. Если в подобные системы добавить модуль, реализующий так называемую настройку на диктора, который перед началом основной работы выполнит анализ особенностей речи говорящего, это позволит значительно снизить вероятность ошибки системы, при этом практически не затронув её производительность.

В связи с изложенным, возникла идея создания программной системы, позволяющей определить пол диктора на основе физиологических характеристик речи, что станет первым шагом к решению проблемы неустойчивости систем распознавания, обусловленной отсутствием привязки речи к типу голоса диктора.

На данный момент существует несколько методов, с помощью которых решается аналогичная задача, к примеру, метод Парзена - распознавание пола диктора в пространстве параметров модели голосового источника, найденных путем решения обратной задачи; метод гауссовых смесей, основанный на моделировании плотности распределения вектора акустических признаков голоса взвешенной суммой нескольких гауссовских распределений; метод, основанный на решении обратной задачи относительно динамики площади голосовой щели и формы импульса объемной скорости потока через голосовую щель.

В данной работе для решения задачи идентификации пола диктора были использованы методы теории активного восприятия.

Алгоритм, основанный на применении теории активного восприятия, включает три этапа:

1) этап предварительной обработки сигнала (формирование исходного описания сигнала);
2) формирование системы признаков сигнала;
3) этап классификации сигнала на основе системы признаков.

Третий этап – классификация, выполненная на основе полученной системы признаков, осуществляется путем использования существующих классификаторов, вариации которых могут влиять на точность конечных результатов. В данной работе для получения срав-

нительных данных будут использоваться два классификатора: svm, основанный на методе опорных векторов, и knn, основанный на методе *k* ближайших соседей.

Метод опорных векторов был упомянут ранее, что касается knn, его идея заключается в вычислении некоторого заданного количества *k* ближайших соседей этого объекта (объектов, классы которых уже известны), тогда класс исходного объекта определяется тем, какой класс наиболее многочислен среди этих соседей.

В ходе тестирования программной системы, созданной на основе перечисленных методов и алгоритмов, были получены следующие результаты:

1. Для классификатора svm при значении параметров cost = 5000 (цена нарушения ограничений) и gamma = 0,0005 (параметр, определяющий насколько мало влияние одного тренировочного объекта на результат классифицирования) и при проведении тестирования на массиве из 32 тестовых записей результат оказался следующим: для класса Female было угадано 13 из 16 записей, для класса Male - 16 из 16. В общем получается, что количество успешно распознанных записей - 29 из 32, а это 90,6%.

2. Для классификатора knn при значении $k = 7$ при проведении тестирования на массиве из 32 тестовых записей результат следующий: для класса Female было угадано 12 из 16 записей, для класса Male - 13 из 16. В общем получается, что количество угаданных записей - 26 из 32, а это 81,25% (результат хуже, чем у svm).

Для сравнения, распознавание пола диктора с помощью метода Парзена даёт результат до 97%, распознавание пола на основе решения обратной задачи относительно динамики площади голосовой щели и модели одномерного потока через голосовую щель даёт результат до 94,7% точности для распознавания мужского голоса, и до 97,6% - для распознавания женского; использование метода на основе моделирования акустических параметров голоса гауссовыми смесями даёт точность до 91%. Таким образом разработанный алгоритм при использовании для классификации метода опорных векторов даёт сравнительный с другими системами результат при использовании относительно небольшого количества обучающих записей.

Заключение

В приведенных работах рассмотрены методы анализа информации, заключенной в звуковых записях. Решение задач выполнялось с позиции теории активного восприятия. Методы были реализованы программно и протестированы.

Полученные результаты вычислительных экспериментов подтверждают эффективность предложенных методов. Созданные системы могут быть использованы во многих сферах применения, также они имеют потенциал для дальнейшего развития.

Библиографический список

1. **Nwe, T. L.** Speech emotion recognition using hidden Markov models / T. L. Nwe, S. W.Foo, L. C. De Silva // Speech communication. – 2003. – V. 41. – № 4. – P. 603–623.
2. Perez-Meana H. (ed.). Advances in Audio and Speech Signal Processing: Technologies and Applications: Technologies and Applications // Igi Global, 2007, Ch. 13. – P. 374.
3. **Utrobin, V. A.** Physical interpretation of the elements of image algebra // J. Advances in Physical Sciences. – 2004. – № 47. – P. 1017–1032.
4. Как это работает? Рекомендации в Яндекс. Музыке. URL: https://yandex.ru/blog/company/92883
5. Машина опорных векторов. URL: http://www.machinelearning.ru/wiki/index.php?title=SVM
6. **Гай, В.Е.** Информационный подход к описанию звукового сигнала // Труды МФТИ. – 2014. – Т. 6. – № 2. – С. 167–173.
7. **Гай, В.Е.** Оценка эмоционального состояния человека по голосу с позиций теории активного восприятия / В.Е. Гай [и др.] // Системы управления и информационные технологии. – 2015. – №1.1 (59). – С. 118–122.
8. **Вапник, В.Н.** Восстановление зависимостей по эмпирическим данным / В.Н. Вапник. – М.: Наука, 1979. – 448 с.

9. **Hastie, T.** The Elements of Statistical Learning, 2nd edition / T. Hastie, R. Tibshirani, J. Friedman // Springer. – 2009. – P. 533.
10. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян [и др.]. – М.: Финансы и статистика, 1989. – 198 с.

Дата поступления
в редакцию 01.06.2017

V. E. Gai, I. A. Presnyakov, M. I. Arabaji,
M. O. Derbasov, I. V. Polyakov, E. N. Vikulova

THE ANALYSIS OF AUDIO RECORDINGS FROM THE STANDPOINT OF THE THEORY OF ACTIVE PERCEPTION

Nizhny Novgorod state technical university n. a. R. E. Alekseev

Purpose: New approaches to the analysis of sound signals from the standpoint of the theory of active perception are offered. They allow increasing the accuracy and reducing the computational complexity of methods of pattern recognition.
Metodology: Using the U-transformation and the algebra of groups from the theory of active perception, allows achieving the objectives. The work offers methods for solving the task of assessing the sex of speaker by voice and for solving the task of searching for similar music.
Findings/Research implications: The methods described in the work can find practical application in security systems and in various advisory systems.
Value: The problems of automation of the analysis of sound recordings are considered. The results of computational experiments indicate the effectiveness of the proposed approaches.

Key words: pattern recognition, machine learning, theory of active perception, audio recordings, sound signals, speaker identification, system of recommendations.