АЛГОРИТМ ИДЕНТИФИКАЦИИ ПОЛА ДИКТОРА ПО ГОЛОСУ НА ОСНОВЕ ТЕОРИИ АКТИВНОГО ВОСПРИЯТИЯ

© 2016 г. М.И. АРАБАДЖИ

Нижегородский государственный технический университет им. Р.Е. Алексеева
e-mail: semikova-maria@mail.ru

Задача идентификации пола диктора по голосу является одной 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 - как тех, которые выполняют перевод речи в текст, так и тех, которые направлены исключительно на идентификацию личности диктора. Применение систем идентификации актуально в таких сферах, как криминалистика (идентификация личности преступника на основе записи голоса), сфера безопасности (организация доступа персонала в определённые помещения на основе голосовых данных), сфера разработок голосового управления техникой (так называемые системы "умного дома"). Помимо задачи определения пола диктора в задачу идентификации личности входят так же проблемы оценки возраста диктора, его эмоционального состояния, среднего тона голоса, и так далее. 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 выше и возникла идея создания программной системы, которая позволит 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.

Таким образом, целью данной работы является создание программной системы распознавания пола диктора по голосу, а задача разрабатываемой системы заключается в том, чтобы подготовить почву для дальнейших разработок в этой области, и получить возможность применять разработанную технологию при дальнейшей работе над проблемой идентификации личности 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.

Основной алгоритм, основанный на применении теории активного восприятия включает в себя три этапа:

–этап предварительной обработки сигнала (формирование исходного описания сигнала);

– формирование системы признаков сигнала;

– этап классификации сигнала на основе системы признаков.

Целью первого этапа – предварительной обработки сигнала является представление сигнала в удобной форме для последующего анализа, а именно - формирование его спектрального представления с помощью *U*-преобразования (поэтапной фильтрации, интегрирования и пространственного дифференцирования сигнала).

В рамках теории активного восприятия входящий сигнал рассматривается как системное преобразование. Для обнаружения его структурных элементов используется операция интегрирования, а для обнаружения связей между этими элементами - пространственное дифференцирование. Вместе эти два преобразования и называются *U*-преобразованием (1):

$$U = \mathrm{d}\int \tag{1}$$

Следующий этап алгоритма - формирование системы признаков сигнала. Делается это следующим образом: вся полученная масса измерений (отсчётов) разбивается на 16 частей, а затем амплитуды отсчётов, относящиеся к каждой части, суммируются (рис. 1):

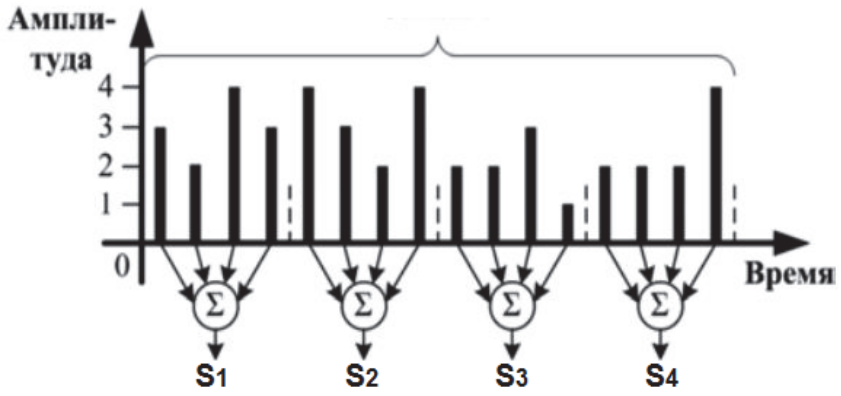

Рис. 1. Сложение амплитуд отсчётов сигнала.

Затем к получившемуся массиву сумм *S* применяются базисные одномерные фильтры-покрытия, позволяющие получить набор спектральных коэффициентов. Исследования показали, что для целей анализа звукового сигнала необходимо использовать не менее 16 фильтров. Фильтры применяются следующим образом: выполняется операция сложения над суммами отсчётов, причём те суммы, которые попали в положительную область фильтра, берутся с положительным знаком, а те, которые попали в отрицательную область - с отрицательным. Таким образом, для каждого фильтра мы получаем некоторое число, называющееся спектральным коэффициентом. С помощью этих коэффициентов формируется огрублённое представление спектрального представления сигнала для приведения значений спектральных коэффициентов сигнала к конечному, заранее заданному набору амплитуд с заданной степенью огрубления - *k*. После получения огрублённого представления сигнала определяются ключевые сегменты - участки сигнала, на которых максимален отклик небольшого количества операторов. Поиск ключевых сегментов используется для того, чтобы можно было уменьшить количество анализируемых сегментов за счёт отфильтровывания шумовых сегментов, пауз, и так далее.

Затем формируется описание сигнала с помощью спектрально-корреляционного анализа с использованием так называемых полных и замкнутых групп. Множества этих групп формируются на основе 16 бинарных операторов $\{V_i\}$. Операторы вычисляются по фильтрам таким образом: для значения фильтра прини-

мается, что (+1→ 1), (-1→ 0), и тогда, при условии допустимости для операторов операций объединения и пересечения, имеем алгебру групп (2):

$$A_v =< \{V_i\}: +, > \quad (2)$$

В алгебре A_vсуществуют полные группы, образованные на трёх операторах и позволяющие выявить корреляционные связи между операторами, и замкнутые группы, образованные на четырёх операторах, и позволяющие выявить корреляционные связи между полными группами.

Для получения системы признаков сигнала необходимо сформировать множества полных и замкнутых групп, затем для каждого ключевого сегмента вычислить замкнутые группы и отобрать некоторое количество *N* первых по массе групп. И, наконец, выполнить отбор устойчивых сегментов сигнала (тех сегментов, для которых разница между максимальной и минимальной массами групп, входящих в описание сигнала, максимальна).

Последний этап обработки звукового сигнала - его классификация, выполненная на основе полученной системы признаков. Данный этап осуществляется путем использования существующих классификаторов, вариации которых могут влиять на точность конечных результатов. В нашей работе для получения сравнительных данных будут использоваться два классификатора: svm, основанный на методе опорных векторов, и knn, основанный на методе *k* ближайших соседей.

Классификатор, как правило, включает в себя тренировочные и тестовые наборы, состоящие из экземпляров данных. Каждый экземпляр в наборе содержит одно целевое значение (класс, в нашем случае этот класс определяет пол диктора), и несколько атрибутов (признаков). Цель классификатора заключается в создании модели, способной прогнозировать целевые значения экземпляров данных в наборе тестирования, для которых известны только атрибуты

Идеей классификатора svm является поиск оптимальной гиперплоскости, то есть такого линейного классификатора, который обеспечит максимальную дистанцию между собой и ближайшими примерами объектов из каждого класса, а идея классификатора knn заключается в вычислении некоторого заданного количества *k* ближайших соседей этого объекта (объектов, классы которых уже известны), и тогда класс исходного объекта определяется тем, какой класс наиболее многочислен среди этих соседей.

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

- 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 угаданных записей - 29 из 32, а это 90.6%, очень хороший результат.
- 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 значительно 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, видно, что разработанный алгоритм при использовании для классификации метода опорных векторов даёт нормальный результат по сравнению с другими

системами, особенно с учётом использования относительно небольшого количества обучающих записей.

В заключение заметим, что существует два пути дальнейшего применения описанного алгоритма в программных системах: первый путь подразумевает применение системы идентификации пола диктора по голосу в качестве дополнительного модуля к системам, работающим над распознаванием речи (переводом её в текст). Это предложение обусловлено пониманием того, что большинство систем распознавания речи не уделяют должного внимания физиологическим особенностям голоса, таким как общий тон, интонационная окраска, особенностям, связанным с полом и возрастом говорящего, а тем временем, учёт этих параметров на моменте подготовки к распознаванию поможет значительно снизить количество ошибок в системе.

Второй путь возможного развития – разработка системы, основанной на данном алгоритме, не в качестве прикладного модуля, а в качестве самостоятельной, направленной на решение задачи идентификации человека по голосу, то есть определения максимального набора особенностей голоса, позволяющего с некоторой точностью идентифицировать личность говорящего.

СПИСОК ЛИТЕРАТУРЫ

1. *Гай В.Е.* Информационный подход к описанию звукового сигнала // Труды МФТИ, 2014 г., Том 6, № 2, с. 167-173.
2. *Гай В.Е., Утробин В.А., Родионов П.А., Дербасов М.О.* Оценка эмоционального состояния человека по голосу с позиций теории активного восприятия. // Системы управления и информационные технологии, №1.1 (59), 2015 г., с. 118-122.
3. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным // Наука, Москва, 1979 г., с. 448.
4. *Hastie, T., Tibshirani, R., Friedman, J.* The Elements of Statistical Learning, 2nd edition // Springer, 2009, p. 533.
5. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: классификация и снижение размерности // Финансы и статистика, Москва, 1989 г.