КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.9

РАСПОЗНАВАНИЕ ЗВУКОВЫХ ОБРАЗОВ НА ОСНОВЕ АНАЛИЗА ОТКЛИКА СИСТЕМЫ ОСЦИЛЛЯТОРОВ

Д. В. Болотнов, С. А. Запрягаев

Воронежский государственный университет

Поступила в редакцию 01.02.2012 г.

Аннотация. В статье рассматривается проблема выделения отличительных признаков сигнала на основе анализа отклика системы связанных осцилляторов на акустический речевой сигнал.
Ключевые слова: распознавание речи, дискретное преобразование Фурье, система осцилляторов.
Annotation. The paper contains an overview of the most commonly used approaches to solving the problem of speech recognition. It's considered a question of extraction acoustic features for resolving speech recognition problem on the base of oscillators reactions under the acoustic speech signals.
Keywords: speech recognition, Discrete Fourier Transform, system of the oscillators.

Распознавание несложных звуковых образов является одним из подходов при решении задачи распознавания человеческой речи. Речь является основной формой общения людей. Речевой звук представляет собой колебание плотности воздуха, созданное человеком в горловом тракте, воздействующее на барабанные перепонки человеческого уха и выполняющие роль промежуточного звена при анализе звука. В технических устройствах в роли аналога уха выступает микрофон, который в сочетании с аналогово-цифровым преобразователем позволяет получить дискретный набор данных, характеризующих амплитуду колебания во времени. Анализ данного набора данных и преобразования речи в текстовое представление рассматривается как один из вариантов распознавания речи.

Для обработки речи достаточно частоты дискретизации, которая лежит в диапазоне от 8 кГц до 16 кГц. Принципиальным отличием технического подхода распознавания от процесса, происходящего в организме человека, является то обстоятельство, что анализируется одномерный оцифрованный сигнал, в то время как ухо человека анализирует объемный сигнал или, по крайней мере, сигнал колебания поверхности барабанной перепонки. В настоящей работе процесс приема звукового колебания моделируется системой связанных осцилляторов. Соответственно, распознавание осуществляется на основе анализа отклика такой системы под воздействием звука.

Хорошо известно, что одна из задач распознавания речи заключается в удалении посредника при общении человека с компьютером. Это позволит получить и новые методы управления техникой непосредственно человеческим голосом, упростит и ускорит ввод информации. Практически реализация такого интерфейса является более простой задачей, так как оперирует в этом случае с ограниченным набором звуковых образов, но, несомненно, является необходимым первым шагом на пути решения полной задачи распознавания речи. Проблема распознавания речи достаточно интенсивно исследована и по многим направлениям решена для ряда европейских, часто используемых языков. Дело с русским языком обстоит иначе. Несмотря на то, что русский язык занимает 8-ое место в рейтинге языков по численности говорящих на этом языке [1], системы автоматического распознавания русской слитной русской речи развиты незначительно. В связи с этим проведение исследования методов распознавания звуковых образов и речи русского языка является достаточно актуальной и востребованной задачей.

1. ОСНОВНЫЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЯ В ОБЛАСТИ АНАЛИЗА РЕЧИ

Основными современными направлениями исследований в области различных форм анализа и синтеза речи являются [2]:

• *Распознавание речи* – процесс преобразования речевого потока в расшифрованную оцифрованную информацию, используемую по назначению (текст, сигналы управления и т.п.).

• *Выделение ключевых слов* – процесс идентификации звуковых образцов по заранее заданному шаблону.

• *Определение языка сообщения* – процесс поддержания организации работ в распределенных системах управления.

• *Идентификация диктора* – процесс определения личности в системах безопасности.

• *Синтез речи* – процесс обратный процессу распознаванию речи, используемый для озвучивания текстового материала.

В общем случае, по назначению, системы по распознаванию речи можно разделить на три группы: системы приёма команд, системы «речь-текст» и системы преобразования «речь-речь» [3]. При этом *системы приёма команд* – это системы распознавания речи, классифицирующие ограниченное число команд, предназначаемые для управления устройствами. Системы *«речь-текст»* – преобразуют звуковой сигнал в текстовый формат или наоборот. Система преобразования *«речь-речь»* – модифицирует входящий звуковой сигнал, преобразовывая его в заданный звуковой формат.

Все системы распознавания речи можно разделить на две группы: системы, зависимые и независимые от диктора. В случае зависимости распознающей системы от диктора, она настраивается на голос конкретного диктора в процессе обучения (настройка системы). Недостаток такой системы заключается в том, что для работы с голосом другого диктора она должна быть полностью перенастроена (переобучена), что в некоторых случаях может оказаться невозможным. Звуковой образ в таких системах хранится в виде эталонного сигнала, который при работе сравнивается с сигналом, который необходимо распознать. Такие системы хорошо показали себя на фиксированном наборе распознаваемых голосовых команд. В среднем, число таких команд ограничивается тридцатью. При попытке перехода к распознаванию большего количества команд возникает проблема эталонного слова, так как для работы этого метода требуется многократно записать все распознаваемые слова. На первых этапах проектирования систем распознавания речи использовались только такие методы.

Следующим этапом в развитии систем распознавания речи стали системы, работающие независимо от голоса диктора. Такие системы используют алгоритмы распознавания, которые могут работать с любыми голосами, без каких-либо ограничений. Система не требует переобучения и перенастройки на новый голос. Чтобы сделать возможным реализацию этих новых алгоритмов и учёт вариативности речи, потребовалось использование дополнительных методов, основанных на скрытых Марковских моделях (СММ) и искусственных нейронных сетях. В этих случаях, вместо эталона на отдельное слово, потребовалось создание эталонов на части звуков, из которых собираются слова. Такие звуковые части называют акустическими моделями, которые создаются путём анализа большого числа записей различных голосов.

В целом, современные системы распознавания речи реализуют два кардинально различных подхода к решению проблемы распознавания [2]:

• *Распознавание голосовых меток* – представляет собой распознавание фрагментов по заранее записанным эталонам. Данный подход используется, в основном, в несложных системах для распознавания заранее записанных слов.

• *Распознавание лексических элементов* – выделение из слитной речи фонем и аллофонов, как простейших лексических элементов. Данный подход эффективен для разработки систем, считывающих слитную речь в текст, поскольку в нём реализуется постепенное преобразование произнесённых звуков в текст.

По типу *структурной единицы* непрерывную речь делят на: фонемы, морфемы, аллофоны, дифоны, трифоны и слово.

• *Фонема* – минимальная единица звука языка. Фонема не имеет самостоятельного смысла, не имеет грамматического, лексического значения и используются для определения более крупных структурных единиц (например, таких как морфемы или слова) [4]. Замена одной фонемы на другую приводит к другому слову, ровно как и изменяя порядок следования фонем возникает иное слово. Фонема выступает в роли абстрактной единицы языка и соответствует определённому звуку речи. Большинство современных систем распознавания, работающих с фонемным распознаванием, используют скрытые Марковские модели. В ка-

честве аппарата принятия решений используется результат работы нейронной сети, которая передаёт апостериорные вероятности фонем.

• *Морфема* – наименьшая языковая единица, обладающая значением. При делении морфем на части, происходит выделение её составляющих, – фонем. В большинстве лингвистических теорий, морфемы определяются как абстрактные языковые единицы. Выделяют также, так называемые, морфы, определяемые конкретной реализацией морфемы в тексте [4]. При этом морфы, определяющие одну и ту же морфему, могут иметь отличный фонетический облик, что зависит от окружения внутри текущей словоформы. Совокупность морфов в рамках одной морфемы, которые имеют один и тот же фонемный состав, – называют *алломорфом*.

Морфемный анализ используется в качестве дополнения к основному модулю принятия решений при распознавании слитной русской речи. Модель распознавания, основанная на морфемном уровне обработки сигнала, позволяет сократить словарь языковых единиц и ускорить работу системы распознавания.

• *Аллофоны* – реализация фонемы, её разновидность, определяемая текущим окружением словоформы. Аллофон является определённым речевым звуком, в отличии от, например, фонемы, которая являлась абстрактным понятием. Диапазон аллофонов в рамках одной фонемы довольно широк, но человек всегда способен их распознать.

2. МЕТОДЫ РАСПОЗНАВАНИЕ СЛИТНОЙ РЕЧИ

При распознавании слитной речи поэтапно осуществляется несколько уровней обработки сигнала от предварительной его обработки, выделения лексических элементов речи, выделения морфем и слогов до выделения слов, предложений и целых сообщений. Схема работы алгоритма, в случае анализа слитной речи может быть представлена в виде последовательности блоков трех основных совокупностей операций – блок ввода исходных данных, блок распознавания и блок анализа результатов распознавания

Блок ввода исходных данных является вспомогательным блоком имеющим своей задачей подготовку исходного сигнала для дальнейшего анализа. Этот блок может выполнять ряд вспомогательных задач, таких как очистку сигнала от шума, нормировку сигнала, разбиение сигнала на заранее определенные единицы и т.п.

Выделение из анализируемого сигнала набора отличительных признаков, по которым можно произвести классификацию сигнала, осуществляется в блоке распознавания. Существует несколько подходов по выделению из сигнала таких отличительных признаков. Наиболее широко используемым является спектральный анализ сигнала на основе преобразования Фурье. Исследование в области спектрального анализа привели к созданию группы схожих алгоритмов нахождения отличительных признаков сигнала [6]. Такие алгоритмы подразумевают предварительную обработку сигнала, использование быстрого преобразования Фурье (БПФ) от окна сигнала длиной 10–30 мс, построение кепстральных коэффициентов со шкалой Mel или Bark [7].

В системах распознавания речи преимущественно используется так называемое дискретное преобразование Фурье (ДПФ).:

$$X_k = \sum_{n=0}^{N-1} x_n \exp\left(-i\frac{2\pi}{N}k \cdot n\right),$$
$$k = 0,\ldots,N-1,$$

где N – общее количество отсчётов сигнала, x_n –дискретные значения исходного сигнала, X_k – комплексные амплитуду ДПФ.

Одним из самых распространённых подходов к выделению признаков сигнала является спектральный анализ с выделением кепстральных коэффициентов [8, 12]. Кепстр, в общих словах, это спектр логарифма спектра исходного сигнала. Основателем данного подхода является известный советский математик Колмогоров А. Н. Кепстральное преобразование используется главным образом, потому что информация после этого преобразования позволяет максимально исключить влияние речевого аппарата, влияние звуковых резонирующих полостей речевого тракта, сосредоточившись исключительно на речевом сигнале. Формула перевода частотной шкалы (f – герцы) в шкалу Мела имеет вид:

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) =$$
$$= 1127\ln\left(1+\frac{f}{700}\right).$$

Мел-частотный кепстральный анализ – это представление, формируемое спектром от лога-

рифма спектра, проецируемое на нелинейную частотную шкалу Мела. Отличие обычного кепстра от кепстра по частоте Мела заключается в том, что частотная шкала Мела аппроксимирует распределение кепстральных коэффициентов ближе к человеческому слуху

В общем случае Фурье анализ обладает рядом недостатков, таким как, например, потеря информации о временных характеристиках сигналов. Это происходит в следствии использования алгоритмов с частотно-временной локализацией, которые возникают при работе с окнами данных (фреймы, кадры). В современных системах распознавания речи всё чаще можно встретить использование вейвлет-преобразований. В случае с вейвлет-преобразованием, исходный сигнал раскладывается по базисному набору функций, характеризующих частоту и время, поэтому в данном случае анализ может быть проведён с учётом физического и временного пространства переменных.

Кроме перечисленных методов получил определенное развитие метод кодирования коэффициентами линейного предсказания (LPC – Linear Predictive Coefficients). При кодировании коэффициентами линейного предсказания, моделируются параметры речи, передающиеся вместо отсчётов. Данная модель работает с блоками отсчётов, для каждого из которых вычисляется частота основного тона, амплитуда и другие параметры. Алгоритм LPC определяет коэффициенты предсказывающего фильтра путём минимизации среднего квадрата ошибки предсказания.

3. АНАЛИЗ ОТКЛИКА СИСТЕМЫ ОСЦИЛЛЯТОРОВ

Данное исследование представляет собой описание метода распознавания речевых элементов русского языка и результат его сравнения с другими, широко используемыми методами распознавания на основе анализа отклика системы осцилляторов на звуковой сигнал $s(t)$ (или совокупность сигналов $s_k(t),\ k = 1,2,\ldots,n$) для получения отличительных признаков сигнала. Схематично модель можно представить следующим образом (рис. 1).

Разработанная система распознавания состоит из нескольких модулей. Модуль ввода данных отвечает за организацию чтения звуковых файлов. Модуль нормализации данных преобразует исходные данные к единому масштабу, независящему от интенсивности звука. Модуль спектрального анализа обеспечивает необходимые процедуры, связанные с выполнением различных типов Фурье преобразования. Модуль выделения признаков осуществляет построение вектора признаков по заданным образцам и заданным алгоритмам. Последний модуль программной оболочки осуществляет анализ вектора признаков по обученной нейронной сети, которая выдает результат распознавания и оценку ошибки распознавания.

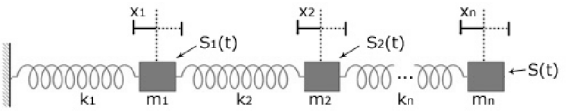

Рис 1. Система связанных осцилляторов, где k_1, k_2 ...— коэффициенты упругости пружин, x_1, x_2 ...— отклонения масс m_1, m_2,.. от положения равновесия. На систему осцилляторов действует сигнал (или набор сигналов), в форме предварительно обработанного звукового сигнала, который приводит гармоническую систему в состояние колебания. В качестве характеристик отклика системы осцилляторов рассматривались траектории и скорости масс системы, а также распределение энергии в системе за время действия возмущающих сигналов. В настоящей работе, метод распознавания, основанный на прямом анализе Фурье образа сигнала, сравнивается с выше приведённым методом анализа отклика специально подобранной системы осцилляторов

Начальный этап обработки сигнала состоит из двух шагов:

1. Считывание wave файла и разбор формата. На данном этапе получается сигнал в виде дискретного набора данных на протяжении определённого отрезка времени, равному времени звучания файла. На файл накладываются определённые технические ограничения: частота дискретизации 11025 KHz, 16 bit pcm, единственная звуковая моно-дорожка. Другие файлы системой не принимаются. Не анализируются стерео-файлы потому, что дорожка другого канала совпадает с первым, что существенно снижает производительность системы при отсутствии преимуществ от стерео-файлов.

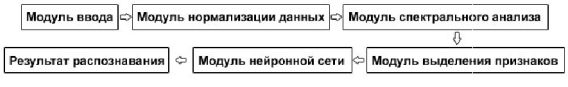

Рис. 2. Блок схема системы распознавания

2. Нормализация сигнала. Прежде, чем данные поступают на последующие модули обработки, они приводятся к общему виду, с которым работает система. По вертикальной оси, диапазон значений нормируется на новый диапазон: 0..10000 для получения равномерно распределённого сигнала. Данное число подобрано экспериментально.

В качестве предварительной обработки сигнала, на следующем этапе производится оконное преобразование сигнала по формуле Хэмминга:

$$\omega(n) = 0.53836 - 0.46164\left(\frac{2\pi n}{N-1}\right)$$

где n — текущее значение сигнала, N — общее количество точек.

После того как сигнал подготовлен к работе, происходит выделение отличительных признаков. Получение отличительных признаков заключается в формировании огибающей кривой полученной от преобразования Фурье (стандартный метод). Из спектра сигнала производится деление спектра на 15 равных областей и вычисление средних значений в каждой из этих областей. Таким образом, получается огибающая спектра из 15 значений, которые формирует вектор отличительных признаков, который и подается на модуль обучения и распознавания.

Система обучения построена на взаимодействии с нейронной сетью. Структура НС состоит из одного скрытого слоя, на котором изначально размещён один нейрон [9]. Тип сети — многослойная сеть с обратным распространением ошибки [10]. Для построения нейронной сети используется программная библиотека FANN.

Эксперимент по сравнению методов распознавания строился на трех вариантах задания вектора признаков:

1. Вектор признаков формируется из огибающей спектра Фурье [11] звукового сигнала;

2. Вектор признаков формируется, из огибающих спектра Фурье сигнала и огибающих отклика системы осцилляторов. В простейшем случае характеристики отклика состоят из Фурье компонент отклонения от положения равновесия масс системы и Фурье компонент скоростей выбранных масс.

3. Вектор признаков формируется только на основе анализа данных огибающих отклика системы осцилляторов под действием звукового сигнала за период действия сигнала.

На первом этапе проверки систем, распознавание проведено трёмя вышеприведёнными вариантами распознавания на примере 21 буквы русского алфавита: а, б, в, г, д, е, ж, и, к, л, м, н, о, п, р, с, т, у, ф, х, ш. На каждую букву для обучения отводилось 15 образцов и 5 образцов для проверки. Таким образом, необходимо распознать 105 образцов (21 буква по 5 образцов на каждую). Результаты распознавания букв русского алфавита на системе из двух связанных осцилляторов с заданными параметрами ($k_1 = 2000$, $k_2 = 1000$, $m_1 = 0.002$, $m_2 = 0.001$) выглядят следующим образом:

1. Метод Фурье: правильно распознано: 81 из 105.

2. Метод Фурье плюс система осцилляторов: 85 из 105.

3. Система осцилляторов: 90 из 105.

Таким образом, метод выделения отличительных признаков на основе анализа отклика рассмотренной модели может оказаться более эффективным методом для решения проблемы распознавания звуковых образов, чем стандартный метод метод фурье анализа сигнала. Существенным элементом метода является выбор параметров системы осцилляторов.

На следующем этапе система была апробирована на распознавании слогов. В качестве эксперимента были использованы слоги: ма, ми, ша, ши, на, ни, им, ин, иш. На вход системы обучения было подано 135 образцов (15 образцов на каждый из слогов) и 45 образцов из другой выборки для проверки обучения. После запуска системы обучения и распознавания были получены следующие результаты:

1. Метод Фурье: 38 правильно распознанных из 45.

2. Метод Фурье + система осцилляторов: 38 правильно распознанных из 45.

3. Система осцилляторов: 38 правильно распознанных из 45.

В данном случае, преимуществ подхода с использованием осцилляторов получено не было. Три результата получились идентичными по степени распознавания, хотя ошибки система допускала в разных слогах. Это может быть следствием того, что двух осциллирующих элементов не хватает для достоверной классификации вышеперечисленных слогов.

Проведенный анализ позволяет сделать вывод об эффективности подхода по выделению отличительных признаков на основе анализа

поведения осцилляторной системы, а полученные преимущества в распознавании изолированных можно использовать в фонемном подходе к распознаванию.

СПИСОК ЛИТЕРАТУРЫ

1. *Encarta: Languages Spoken by More Than 10 Million People* // http://encarta.msn.com/

2. *Фролов А.В., Фролов Г.В.* Синтез и распознавание речи. Современные решения, 1991–2008, Библиотека Братьев Фроловых, www.frolov-lib.ru

3. *Леонович А.А.* Современные технологии распознавания речи.

4. *Карпов А. А.* Модели и программная реализация распознавания русской речи на основе морфемного анализа / Санкт-Петербургский институт информатики и автоматизации Российской академии наук, 2007.

5. *Страуструп Б.* Язык программирования C++ / Б. Страуструп // М.:БИНОМ, 2001. – 1099 с.

6. *Сапожков М.А.* Речевои сигнал в кибернетике и связи / М.А. Сапожков; – М.: Связьиздат, 1963. – 452 с.

7. Распознавание слуховых образов. / Под ред. Н.Г. Загоруйко. – Новосибирск: «Наука», 1970. – 340 с.

8. *Strom N.* Continuous Speech Recognition in the W AXHOLM Dialogue System / N. Strom // Stockholm QPSR, 1996. – pp. 67–95.

9. *Болотнов Д.В., Запрягаев С.А.* Распознавание гласных звуков русского алфавита. – Информатика: проблемы, методология, технологии : материалы XI междунар. науч.-метод. Конф., 2011

10. *Хайкин С.* Нейронные сети. Москва. Вильямс, 2005. – 1104с.

11. *Automatic Speech Recognition.* Massachusetts Institute Of Technology (www.ocw.mit.edu), 2003.

12. *Коновалов А. Ю., Запрягаев С.А.* Распознавания речевых сигналов. – Воронежский государственный университет, 2009.

Болотнов Денис Владимирович – аспирант каф. цифровых технологий Воронежского государственного университета. E-mail: denis1503@gmail.com; Tel. 8 903 030 08 42

Запрягаев Сергей Александрович – д. ф.-м. н, проф. каф. цифровых технологий Воронежского государственного университета. Тел.(4732) 208-257. E-mail: zsa@main.vsu.ru

Bolotnov D.V. – Post-graduate student of the dept. of digital technologies Voronezh State University. E-mail: denis1503@gmail.com; Tel. 8 903 030 08 42

Zapryagaev S. A. – Doctor of Physics-math. Sciences, Professor of the dept. of digital technologies Voronezh State University. Tel.(4732) 208-257. E-mail:zsa@main.vsu.ru