УДК 004.522

К. К. Гладышев, Е. А. Шульгин

СИСТЕМА АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧЕВЫХ КОМАНД

Приводится описание системы автоматического распознавания речевых команд, состоящей из нескольких уровней: оцифровки сигнала; выделения первичных признаков речевых сигналов на основе линейных спектральных корней; поиска распознаваемой команды по базе эталонов с использованием методов динамического программирования; семантической коррекции распознанной последовательности символов по орфоэпическому словарю. Разработанная система может быть использована в сфере речевых технологий.

Ключевые слова: *распознавание речи, динамическое программирование, линейное предсказание.*

Полноценная замкнутая система распознавания речи (СРР) состоит из множества взаимосвязанных уровней [1—4]. Общая эффективность работы такой системы непосредственно зависит от качества реализации каждого уровня. Подобное построение системы связано с физиологическими процессами, происходящими при восприятии речи человеком. На выходе слухового аппарата человека формируется набор сигналов, которые в дальнейшем обрабатываются мозгом и преобразуются в последовательности осмысленных речевых единиц.

Обобщенно можно выделить следующие уровни человеческой системы восприятия речи [5]:

— физическое восприятие колебаний звука в ухе с помощью специального органа — улитки;

— преобразование звуковых колебаний в определенную последовательность информативных сигналов и передача их через нейроны в головной мозг;

— трансформация в головном мозге сигнала, полученного от нейронов уха, в дискретную последовательность речевых единиц;

— семантический уровень, связанный со словарем и грамматикой языка.

Соответственно в системе автоматического распознавания речи, основанной на бионическом подходе, можно выделить следующие уровни:

— преобразование голосового сигнала в последовательность значений амплитуд — оцифровка сигнала;

— формирование совокупности векторов информативных признаков сигнала;

— поиск по базе наиболее близкого к распознаваемой речевой единице эталона;

— нечеткий поиск полученной последовательности речевых единиц по ограниченному словарю;

— построение осмысленных предложений и их коррекция в соответствии с вероятностной речевой моделью языка.

Запись и оцифровка звукового сигнала выполняется через микрофон, подключенный к звуковой карте персонального компьютера. Как показано в работе [6], диапазон частот речевого сигнала составляет от 300 до 4000 Гц. Соответственно достаточная частота дискретизации составляет 8 кГц. Количество уровней квантования звуковой карты (16 бит) также вполне достаточно для восприятия человеком оцифрованного сигнала.

Первичную обработку сигнала можно выполнить при необходимости с помощью специальных программных средств, например произвести различного вида фильтрации или нормировки. Далее оцифрованный сигнал в виде массива значений амплитуд передается на следующий уровень СРР.

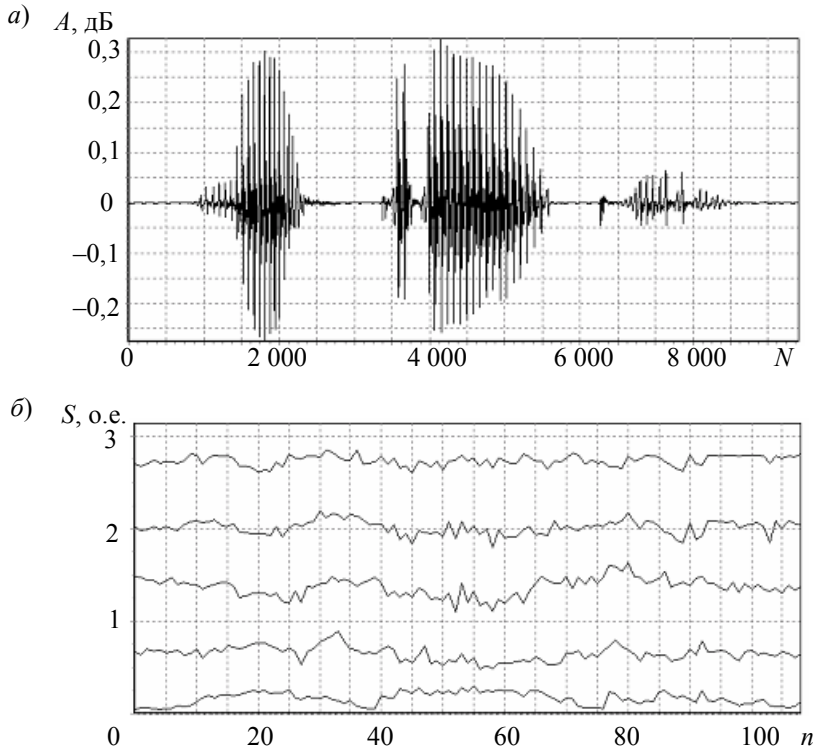
Одним из основных моментов в задачах распознавания речи является выбор метода формирования набора признаков, позволяющих выделить необходимую информацию о речевом сигнале для последующих уровней системы. В современных проектах СРР для выделения информативных признаков используют спектральный, корреляционный и кепстральный анализ, вейвлет-преобразования, линейное предсказание речи.

В системе, разработанной авторами настоящей статьи, для выделения информативных признаков используются линейные спектральные корни (ЛСК). На протяжении многих лет этот аппарат применяется в различных задачах, где необходимо экономно передавать, хранить и достаточно точно воспроизводить информацию о звуковых сигналах [7]. ЛСК позволяют получить информацию о ряде важных параметров речевого сигнала, таких как частота основного тона, ширина полос формантных частот, различные параметры состояния голосового тракта [8]. Кроме того, алгоритмы расчета ЛСК удобны при компьютерной реализации и позволяют производить вычисления в реальном масштабе времени [8].

Согласно новой теории линейных спектральных корней Ланнэ [7] существуют несколько различных вариантов расчета ЛСК. Самым распространенным, хорошо изученным и широко используемым является метод Итакуры, рассматриваемый, например, в работе [9]. Сравнительные исследования нескольких вариантов расчета ЛСК для различных фонем приведены в работе [4]. Так, для рассматриваемого в настоящей статье метода гистограммы ЛСК в пределах одной фонемы не перекрываются, поэтому значения каждого отдельно взятого корня можно эффективно использовать в качестве координаты признакового пространства.

Как показали проведенные исследования, расчет ЛСК целесообразно осуществлять с использованием окна Хэмминга. Для предотвращения потери информации о сигнале окна должны перекрываться. Размер окна выбирается примерно равным периоду основного тона — 10…20 мс — или 80…160 отсчетам при частоте дискретизации 8 кГц.

На рисунке приведен пример осциллограммы слова „настройки“ (*а*) и набор его ЛСК (*б*); здесь A — уровень амплитуды сигнала, N — номер отсчета, S — значение ЛСК, n — номер окна.



Рассчитанная совокупность векторов ЛСК передается на следующий уровень СРР для дальнейшего анализа, где происходит сравнение с эталонами.

В качестве словаря эталонов в рассматриваемой системе взят набор аллофонов (фонем в речевом окружении) русского языка. Фонема является минимальной речевой единицей. Согласно различным фонетическим школам в русском языке насчитываются порядка 40 фонем, из которых 6 — гласных фонем. Однако известно [5], что при плавной речи на звучание каждой фонемы оказывает сильное воздействие ее фонетическое окружение. Например, гласная буква „*о*“ произносится по-разному в словах „в*о*да“ и „водян*о*й“, тем не менее оба этих гласных звука (аллофона) являются разновидностями одной и той же фонемы „*о*“. Так, в современных системах синтеза речи [10] успешно используются именно наборы аллофонов.

По результатам работы [11], где представлены примеры классификации и способы выделения наборов аллофонов, в русском языке выделяются от 300 (минимальный набор) до 2000 (максимальный набор) аллофонов. Для правдоподобного синтеза речи достаточно минимального набора. Так, в разработанной СРР за основу взят упрощенный набор, состоящий из 50 аллофонов. В дальнейшем для повышения точности распознавания словарь эталонов должен быть расширен.

Процедура записи эталонов состоит в том, что диктором начитывается определенная последовательность слов, из которых выделяются необходимые образцы аллофонов. Границы между аллофонами определяются опытным путем. Для каждого аллофона производится расчет информативных признаков — массива линейных спектральных корней. Эти данные записываются в базу эталонов и используются в дальнейшем для сравнения с фрагментом распознаваемой речи.

Для поступающего на вход речевого сигнала для каждого отсчета векторов признаков последовательно производится сравнение с элементами словаря эталонов и выбирается наилучший эталон на основе минимального евклидового расстояния.

На вход распознающей системы могут подаваться и звуки, для которых в базе заведомо нет эталона (например, фонемы других языков или просто нечленораздельные звуки). Поэтому опытным путем установлена максимальная допустимая мера близости к ближайшему найденному эталону. Если эта мера близости превышена, то текущий фрагмент входного сигнала считается нераспознанным.

В результате подобной процедуры распознаваемый сигнал (слово) может быть представлен строкой фонем следующего вида:

<<___ннаааааааа~~~оаооооо~~ииииииииии__>>,

где „_“ — отсутствие сигнала, а „~“ — нераспознанный звук.

Из полученной последовательности удаляются повторяющиеся символы. Обработанная строка фонем используется на следующем уровне СРР, где по словарю производится нечеткий поиск ближайшего слова. Словарь представляет собой таблицу всевозможных слов, воспринимаемых системой, и их фонетических форм. Фактически для полноценной СРР в качестве словаря может использоваться любой орфоэпический словарь русского языка.

Для полученной строки фонем последовательно рассчитывается мера близости L к определенной словарной транскрипции, которая вычисляется по алгоритму нечеткого поиска строк Вагнера и Фишера с использованием метрики Левенштейна [12]. Данный алгоритм позволяет эффективно вычислять значения L для слов различной длины при больших объемах данных. Слово, для которого значение L минимально, считается искомым распознанным.

Ниже приведен пример, демонстрирующий результат поиска ближайшего слова в словаре.

Результат распознавания слова: **н а о и**

Ближайшая транскрипция по словарю: **н а с т р о й к и**

Для устранения грубых ошибок при распознавании введено пороговое максимальное значение L. При превышении данного порога система выдает сообщение об ошибке распознавания и необходимости повторить распознаваемое слово. Пороговое значение определяется экспериментально.

Было проведено тестирование системы в режиме, не зависящем от особенностей речи диктора. Правильно распознано 80 % команд. Объем словаря, позволяющего системе работать в режиме реального времени, составляет 30—40 односложных команд. В настоящее время осуществляется разработка методики по увеличению скорости поиска команд по словарю, что позволит в будущем увеличить его объем при сохранении должного качества распознавания.

СПИСОК ЛИТЕРАТУРЫ

1. *Ямов С. И., Кабак И. С., Курочкин С. Н., Бродин А. Г.* Многоуровневая система распознавания речи // Автоматизация и управление в машиностроении. 1997. № 1.

2. *Лукьяница А. А.* Разработка программы распознавания русской речи для процессора SuperH RISK (Hitachi) / МГУ им. М. В. Ломоносова [Электронный ресурс]: <http://leader.cs.msu.su/~luk/ContinuousSpeech_rus.html>.

3. *Галунов В. И., Галунов Г. В.* Один подход к автоматическому распознаванию речи / Междунар. конф. по компьютерной лингвистике „Диалог — 2000“ [Электронный ресурс]: <http://www.dialog-21.ru/materials/archive.asp?id=6434&y=2000&vol=6078>.

4. *Кисляков С. В.* Разработка и исследование метода распознавания фонем русского языка на основе аппарата линейного предсказания: Автореф. дис. … канд. техн. наук. СПб., 2004.

5. *Чистович Л. А., Венцов А. В.* Физиология речи. Восприятие речи человеком. Л.: Наука, 1976.

6. *Фланаган Д.* Анализ, синтез и воприятие речи М.: Связь, 1968.

7. *Ланнэ А. А.* Новая теория линейных спектральных корней // Тр. 3-й Междунар. конф. „Цифровая обработка сигналов и ее применение“, 29 нояб. — 1 дек. 2000 г., Москва. С.118—125 [Электронный ресурс]: <http://www.dsp.sut.ru/rus/research/publications/download/2000dspa_tom1_30_Lanne.pdf>.

8. *Маркел Дж., Грей А. Х.* Линейное предсказание речи. М.: Связь, 1980.

9. *Ланнэ А. А., Улахович Д. А.* Передача информации о состоянии фильтра-предсказателя с помощью спектральных пар // Радиоэлектроника и связь. 1991. № 1.

10. *Вольская Н., Коваль А., Коваль С.* и др. Синтезатор русской речи по тексту нового поколения // Тр. Междунар. конф. „Диалог — 2005“, 1—6 июня 2005 г., Звенигород [Электронный ресурс]: <http://speechtech.ru/articles/DIALOG_Orator.pdf>.

11. *Лобанов Б. М., Пьорковска Б., Рафалко Я.* и др. Фонетико-акустическая база данных для многоязычного синтеза речи по тексту на славянских языках // Тр. Междунар. конф. „Диалог — 2006“, 31 мая — 4 июня 2006 г., Бекасово. С. 357—364 [Электронный ресурс]: <http://www.dialog-21.ru/dialog2006/materials/html/Lobanov.htm>.

12. *Graham A. Stephen.* Анализ строк / Пер. с англ.; Под ред. *П. Н. Дубнера* // Материалы по математической статистике и программные алгоритмы [Электронный ресурс]: <http://infoscope.ws/string_search/Stephen-92/index.html>.

Сведения об авторах

Константин Константинович Гладышев — аспирант; Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, кафедра цифровой вычислительной техники; E-mail: gladkos@gmail.com

Евгений Александрович Шульгин — д-р техн. наук, профессор; Невский институт языка и культуры, Санкт-Петербург, проректор по информационным технологиям; E-mail: eshu1944@mail.wplus.net

Рекомендована кафедрой цифровой вычислительной техники СПбГУТ

Поступила в редакцию 11.12.07 г.