УДК 004.934.2

ПРИМЕНЕНИЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ПРИ СЕГМЕНТАЦИИ РЕЧИ

[1]Желтов П.В., [1]Семенов В.И., [1]Желтов В.П.

[1]ФГБОУ ВПО «ЧГУ им. И.Н.Ульянова», Чебоксары, Россия (428015, Россия, г. Чебоксары, Московский проспект, 15), e-mail: chnk@mail.ru

В статье рассматриваются существующие методы сегментации речевого сигнала. Среди разнообразных созданных в настоящее время методов можно выделить следующие: метод динамического искажения времени (Dynamic Time Warping, DTW), марковские Модели (Hidden Markov Models, HMM), искусственные нейронные сети (Artificial Neural Networks, ANN), прочие методы на основе DBN (Dynamic Bayesian Networks) и SVM (Support vector machines). Рассматриваются методы формирования слова, принципы, по которым производится отличие некоторых слов. Представлена математическая модель распознавания речи. В отличие от печатного текста или искусственных сигналов естественная речь не допускает простого и однозначного членения на элементы (фонемы, слова, фразы), поскольку эти элементы не имеют явных физических границ. Они вычленяются в сознании слушателя, носителя данного языка, в результате сложного многоуровневого процесса распознавания и понимания речи.

Ключевые слова: вейвлет-преобразование, речь, энергия сегментов фонем, математическая модель, распознавание речи, энергия сегментов, двумерный объект

REVIEW OF EXISTING METHODS OF SEGMENTATION OF SPEECH SIGNALS

[1]Zheltov P.V., [1]Semenov V.I., [1]Zheltov V.P.

[1]Chuvash State University, Cheboksary, Russia (428015, Cheboksary, Russia, Moskovskiy prospect 15), e-mail: chnk@mail.ru

The article deals with the existing methods of speech signal segmentation. Among various existing methods one can point out following: Dynamic Time Warping (DTW), Hidden Markov Models (HMM), Artificial Neural Networks (ANN), other methods based on DBN (Dynamic Bayesian Networks) and SVM (Support vector machines). Are considered methods of forming words, the principles that make the words unlike. In the article is presented the mathematical model of solid speech recognition. Unlike the printed text or artificial signals natural speech does not allow simple and unequivocal partition (to phonemes, words, phrases), as its elements don't have physical boundaries. They are recognized in the mind of the listener – the speaker of the language – as a result of multilevel process and understanding of speech.

Keywords: wavelet-transformation, speech, phonemes segments energy, mathematical models, speech recognition, segments energy, two dimensional object.

Задача сегментации речевого сигнала решается как при создании обучающих баз данных, содержащих фразы с информацией об их сегментации на звуки, так и во время работы систем распознавания речи, основанных на фонемном подходе для выделения из потока речи конкретных звуков. Для сегментации речевого потока применяется большое количество разнообразных алгоритмов, во многих из них система должна быть предварительно обучена.

Среди разнообразных созданных в настоящее время методов можно выделить следующие.

1. Метод динамического искажения времени (Dynamic Time Warping, DTW) – алгоритм, позволяющий вычислить степень схожести речевого фрагмента с существующим в системе эталоном. Подход базируется на работах Р. Беллмана [1] и Т.К. Винцюка [2]. Большое значение имеет тот факт, что одна и та же фраза, произнесенная одним и тем же

диктором, может различаться по длительности произношения, кроме того, различную длительность будут иметь и составляющие фразу (слово) звуки. С учетом данного факта метод подразумевает выравнивание двух речевых сигналов по временной шкале при максимальном совмещении между собой сегментов с одинаковыми звуками. После этого происходит вычисление оценки различия сегментов двух данных произнесений.

Суть метода заключается в поиске такой оптимальной последовательности пар сегментов, которой бы соответствовала минимальная суммарная оценка различия. Визуально работу данного подхода можно представить с помощью матрицы, по горизонтали и вертикали которой отложено время, по оси ординат откладывается фраза-эталон, а по оси абсцисс – фраза, произнесенная для распознавания.

Преимущество данного подхода заключается в простоте установления временного соответствия между проверяемым и эталонным речевым фрагментом, что делает возможным нахождение меры различия между ними.

Основной недостаток данного подхода следует из особенностей его работы, а именно требований по наличию эталонов для всех распознаваемых единиц (для всех дикторов). Данный факт ограничивает область использования метода только случаями распознавания небольшого словаря команд и требует при этом предварительного обучения системы.

2. Марковские модели (Hidden Markov Models, HMM) – статистическая модель, имитирующая работу процесса, похожего на Марковский процесс, с неизвестными параметрами. Метод пользуется большой популярностью благодаря простоте и удобству его применения на практике. Одним из главных недостатков данного подхода является наличие условия о независимости последовательности векторов признаков, вычисленных для речевого фрагмента, что делает невозможным отслеживание закономерностей и взаимных влияний в сигнале на достаточно больших промежутках времени. Однако для компенсации влияния данного фактора были разработаны методы вычисления векторов признаков речевого сигнала, несущих в себе, помимо абсолютных значений характеристик, и значения того, как быстро они изменяются во времени относительно предыдущих фрагментов.

3. Искусственные нейронные сети (Artificial Neural Networks, ANN) – алгоритмы на основе нейросетей отличаются хорошей устойчивостью к наличию посторонних шумов, низкому качеству записи, делают возможным моделирование длительных по времени взаимосвязей в речи. Это позволяет широко использовать при вычислениях параллельное программирование. Среди недостатков нейросетевого подхода выделяют сложность его адаптации для применения в условиях, когда анализируемый сигнал может иметь различную длительность, проблемы выбора начальной структуры и характеристик сети, а также

практическую невозможность извлечения данных о выявленных нейросетью закономерностях.

4. Прочие методы – среди остальных используемых методов выделяются методы на основе DBN (Dynamic Bayesian Networks) и SVM (Support vector machines).

Подводя итог вышеизложенному, можно отметить, что существует набор методов, на базе которых строятся программные средства распознавания речи. Данные методы имеют свои достоинства и недостатки, а некоторые из них, такие как метод DTW, обладают ограниченной областью использования ввиду имеющихся особенностей.

Чтобы отличить некоторые слова, используются дополнительные признаки. Число сегментов, приходящихся на гласные и шипящие буквы, как правило, больше числа сегментов, приходящихся на остальные. Фурье-спектр сегментов вейвлет-преобразования шипящих букв сильно отличается от других букв.

Одной из основных трудностей при распознавании является неопределенная временная организация сигнала. Точность распознавания слов существенно зависит от точности определения границ фонем. Определение границ фонем означает операция целесообразного разбиения речи на фрагменты, т.е. сегментации речи.

Большой интерес представляет применение непрерывного вейвлет-преобразования для сегментирования речи по энергии сегментов.

Для вычисления энергии сегментов фонем используется формула Парсеваля:

$$\int_{-\infty}^{\infty} f^{2}(t)dt = \int_{-\infty}^{\infty} \left|F(\nu)\right|^{2} d\nu .$$

Непрерывное-вейвлет преобразование определяется формулой:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} S(t)\ \psi\left(\frac{t-b}{a}\right)\ dt\ .$$

Для определения границ между гласными и согласными буквами слова вычисляется энергия сегментов функций *W(a,b),* для масштабных коэффициентов 2 и 20, т.е. *W*(2,*b*), *W*(20,*b*). Используя полученные результаты, находим Фурье-спектр сегментов функций *W*(2,*b*), *W*(20,*b*) и речевого сигнала *S*(*t*). Энергия сегментов вычисляются по формуле:

$$E = \sum_{i=1}^{n} F(i).$$

Положительные и отрицательные значения вейвлет-коэффициентов почти одинаковы, поэтому среднее значение вейвлет-коэффициентов в сегменте близко к нулю. На рисунке 1 приведен график сегментов *E*З(*n*) слова *осень*. На рисунке 2 представлена энергия сегментов

*E*1(*n*) функции *W*(1,*b*) слова *осень*. При сравнении рисунков сразу видно различие функций *E*1(*n*), *E*3(*n*).

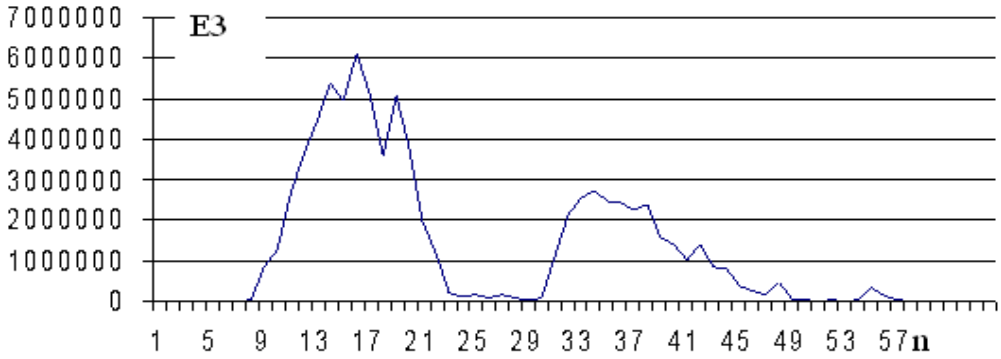

Рис. 1. Энергия сегментов E3(n) слова осень

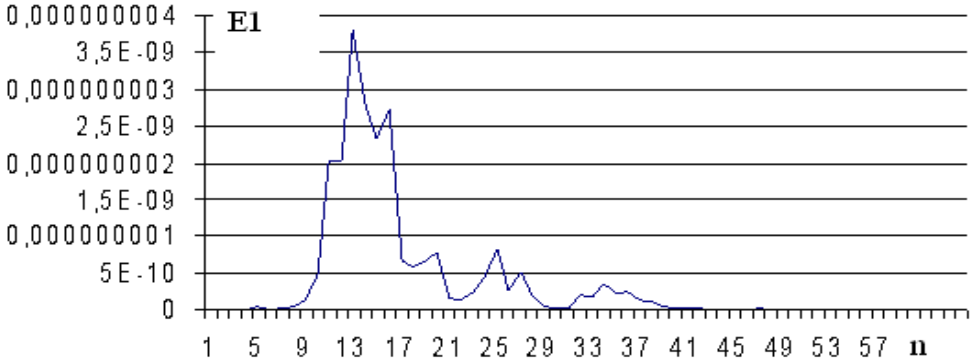

Рис. 2. Энергия сегментов E1(n) функции W(1,b) слова осень

Анализ показывает, что энергия сегментов гласных букв в *W*(1,*b*), *W*(2,*b*) выделяется в виде максимальных пиков. Энергия согласных букв меньше, чем энергия гласных. Энергия сегментов шипящих букв в *E*1(*n*) выделяется в виде максимальных пиков, в *E*2(*n*) и *E*3(*n*) – виде минимумов.

Для увеличения скорости вычисления используем быстрое преобразование Фурье (БПФ), вычисляются коэффициенты тригонометрического ряда $a_1(n), b_1(n)$ функций *EI*(*k*):

$$a_1(n) = \frac{1}{N}\sum_{k=0}^{N-1} EI(k)\cos(\frac{2\pi nk}{N}).$$

$$b_1(n) = \frac{1}{N}\sum_{k=0}^{N-1} EI(k)\sin(\frac{2\pi nk}{N}).$$

Вычисляются коэффициенты тригонометрического ряда $a_2(n), b_2(n)$ вейвлета ψ(*k*) с использованием БПФ:

$$a_2(n) = \frac{1}{N}\sum_{k=0}^{N-1} \psi(k)\cos(\frac{2\pi nk}{N}).$$

$$b_2(n) = \frac{1}{N}\sum_{k=0}^{N-1}\psi(k)\sin(\frac{2\pi nk}{N}).$$

Определим спектр:

$$c_1(n)=a_1(n)\cdot a_2(n)+b_1(n)\cdot b_2(n),$$
$$c_2(n)=b_1(n)\cdot a_2(n)-a_1(n)\cdot b_2(n).$$

Так как *MHAT*-вейвлет четный:

$$c_1(n)=a_1(n)\cdot a_2(n),$$
$$c_2(n)=b_1(n)\cdot a_2(n).$$

Используя обратное преобразование Фурье, получим:

$$WI(4,n) = \sum_{k=0}^{N-1} c(k)\exp(i\frac{2\pi nk}{N}).$$

Математической моделью речевого сигнала при нахождении границ между гласными и согласными звуками речи служит вейвлет-спектр энергии сегментов вейвлет-спектра речевого сигнала. Коэффициент a меняется от 3 до 8. Обозначим их $W1(4,b)$, $W2(4,b)$ и $W3(4,b)$ соответственно. Здесь b меняется от 1 до 256. На рисунке 3 представлен результат ВП функции $E2(n)$ слова *сигнал*. На рисунке 3 положительным значениям функции $W2(4,b)$ соответствуют гласные звуки, а отрицательным значениям — согласные.

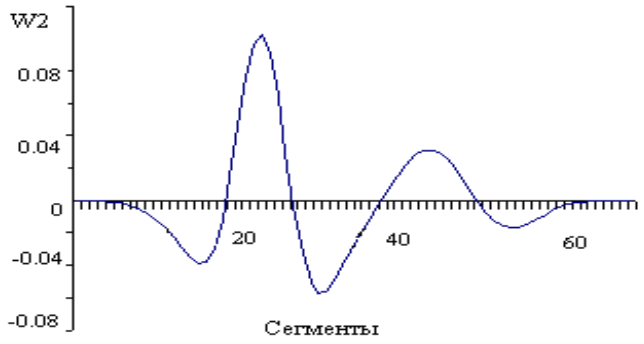

Рис. 3. Вейвлет-спектр W2(4,b) функции E2(n) слова сигнал

Исследования показывают, что гласным буквам соответствует положительное значение в $W1(4,b)$, $W2(4,b)$ и $W3(4,b)$. Шипящим согласным соответствует отрицательное значение в $W2(4,b)$ и $W3(4,b)$. Некоторые шипящие буквы имеют положительное значение в $W1(4,b)$. Для нахождения местоположения гласных букв нормируются энергии $E2(n)$, $E3(n)$ находится их сумма, выполняется вейвлет-преобразование $W4(4,b)$. Таким образом, если слово содержит одну гласную букву, то выделяется один положительный максимум, если две гласные буквы – два положительных максимума и т.д. Каждое слово имеет

определенную структуру. Граница между гласными и согласными буквами или между гласными и шипящими определяются с точностью до 2–3 сегментов.

Для формирования слова подсчитывается количество распознанных фонем *а* в интервале, где выделяются гласные буквы. Аналогично для других гласных букв по отдельности находится количество распознанных букв.

Такие буквы, как *м*, *н*, *л*, имеют почти одинаковые признаки, поэтому для идентификации слов используется словарь, чтобы проверить наличие составленных слов в базе данных слов. Слова, состоящие из трех букв, сравниваются со словами в словаре.

Например, слово *яма* можно в словаре написать в четырех вариантах: *яма*, *амя*, *яла*, *ена* в строковом массиве $x10(i, j)$. Для всех этих сочетаний букв в строковом массиве $y10(i, j)$ сохраняются буквы *я*, *м*, *а*, т.е. для i, $i + 1$, $i + 2$, $i + 3$, потому что русских слов *амя*, *яла*, *ена* нет. Фонемы *а*, *я* имеют почти одинаковые признаки, фонемы *н*, *л*, *м* между собой мало отличаются. Кодирование одного слова *яма* несколькими вариантами увеличивает вероятность распознавания этого слова. Другие слова в словаре хранятся также в нескольких вариантах. Аналогичный алгоритм используется для слов, состоящих из четырех, пяти, шести и более букв. Если слово имеет две гласные буквы, т.е. вейвлет-спектр $W4(4,b)$ состоит из двух положительных максимумов, для идентификации слова применяется сначала алгоритм для шести букв. Вариантов расположения букв несколько. Например, две согласные буквы рядом, две через гласную букву. Потом для пяти, четырех и трех букв.

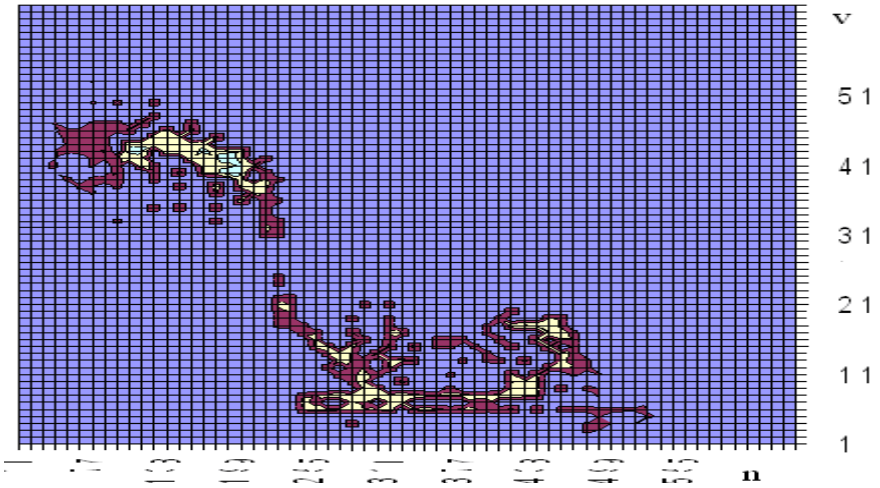

Рис. 4. Фурье-спектр сегментов ВП слова щебень

Базу данных отдельных слов можно использовать для всех людей, потому что структура слова не зависит от различного произношения, тембра и эмоционального состояния человека, а границы между гласными и согласными определяются для всех людей одинаково.

Для того чтобы отличить некоторые слова, используются дополнительные признаки. Число сегментов, приходящихся на гласные и шипящие буквы, как правило, больше числа

сегментов, приходящихся на остальные. Фурье-спектр сегментов ВП шипящих букв сильно отличается от других букв. На рисунке 4 представлен Фурье-спектр сегментов ВП слова *щебень*. На рисунке видно, что фонема *щ* имеет отличный от других спектр. Шипящие фонемы при произнесении содержат больше высокочастотных составляющих в спектре, чем остальные фонемы.

Многомасштабная обработка речевого сигнала выделяет глухие взрывные звуки при большом масштабном множителе, глухие щелевые и аффрикаты – при малом значении масштабного множителя. Гласные фонемы имеют наибольшие значения вейвлет-коэффициентов при средних значениях масштабного множителя и большую длительность по сравнению с другими звуками речи.

Работа выполнена при поддержке РФФИ, проект № 14-07-00143 а.

Список литературы

1. Беллман, Р. Динамическое программирование [Текст] / Р. Беллман. – М.: Иностранная литература, 1960. – 400 с.

2. Винцюк, Т.К. Распознавание слов устной речи методами динамического программирования [Текст] / Т.К. Винцюк. // Кибернетика. – 1968. – № 1. – С. 81–88.

3. Желтов П.В., Семенов В.И., В.П. Желтов. Распознавание слитной речи // Вестник Чувашского университета. № 3. 2012 г. С. 208–210.

4. Желтов П.В., Семенов В.И., В.П. Желтов. Математическая модель распознавания слитной речи // Вестник Чувашского университета. №3, 2012 г. С. 210–212.

5. Семенов В.И., Желтов П.В. Вейвлет-преобразование акустического сигнала / КГТУ им. А.Н. Туполева. Казань, 2008. 102 с.

Рецензенты:

Охоткин Г.П., д.т.н., профессор, заведующий кафедрой автоматики и управления в технических системах ФГБОУ ВПО «Чувашский государственный университет имени И.Н. Ульянова, г. Чебоксары;

Славутский Л.А., д.ф.-м.н., профессор кафедры автоматики и управления в технических системах ФГБОУ ВПО «ЧГУ им. И.Н. Ульянова, г. Чебоксары.