ИНФОРМАТИКА И УПРАВЛЕНИЕ В ТЕХНИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМАХ

УДК 004.932

О. Н. Корелин, Е. Ю. Леонова, П. П. Танонов

ПРИМЕНЕНИЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ И СКРЫТЫХ МАРКОВСКИХ МОДЕЛЕЙ В РАСПОЗНАВАНИИ РЕЧИ

Нижегородский государственный технический университет им. Р.Е. Алексеева

Представлен анализ применения вейвлет-преобразования и скрытых марковских моделей в задаче распознавания речи. Рассмотрены разные варианты конфигурации системы с использованием определённых результатов вейвлет-преобразования и различных объёмов словаря. Для сравнения сигналов использован алгоритм динамической трансформации временной шкалы.

Ключевые слова: распознавание речи, вейвлет-преобразование, dynamic time warping, скрытые марковские модели.

В данной работе опробован алгоритм распознавания изолированных слов, использующий вейвлет-преобразование для цифровой обработки сигнала, а также скрытые марковские модели для оценки результатов распознавания. Данный алгоритм настроен на распознавание изолированных слов, произносимых одним диктором.

Вейвлеты - это функции в виде коротких волн (всплесков) с нулевым интегральным значением и с локализацией по оси независимой переменной (t или x), способных к сдвигу по этой оси и масштабированию (растяжению/сжатию).

При вейвлет-анализе в связи с масштабированием вейвлеты способны выявить различие в характеристиках процесса на различных шкалах, а посредством сдвига можно проанализировать свойства процесса в различных точках на всем исследуемом интервале:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\,\Psi\left(\frac{t-b}{a}\right), \qquad (1)$$

где a – масштаб, малый масштаб соответствует высоким частотам, и наоборот; b – сдвиг, он задаёт положение вейвлета на временной оси.

Для определённых вейвлетов может быть построена масштабирующая функция. Сигнал может быть представлен в виде суперпозиции масштабирующих функций с различными значениями сдвига и масштаба. Получившиеся для различных масштабов сигналы называют ***аппроксимирующими***.

Также сигнал представляют в виде суперпозиции вейвлетов с различными значениями сдвига и масштаба. Получившиеся сигналы называют ***детализирующими***, а образуемую ими трёхмерную плоскость – вейвлет – спектром сигнала. Детализирующие сигналы также могут быть получены как разность аппроксимаций соседних уровней масштаба.

Подобный анализ сигнала называют непрерывным вейвлет-преобразованием (**НВП**). НВП является избыточным, поэтому на практике применяют дискретное вейвлет-преобразование (**ДВП**).

В ДВП при анализе сигналов значения масштаба a и сдвига b дискретизируют.

© Корелин О. Н., Леонова Е. Ю., Танонов П. П., 2016.

$$a=2^m,\ b=k\cdot 2^m. \qquad (2)$$

Сигнал также дискретизируется по времени.

При подобной дискретизации исключается перекрытие носителей вейвлетов, таким образом устраняется избыточность НВП.

Определённые вейвлеты (например, Добеши) и их масштабирующие функции могут быть представлены в виде набора коэффициентов цифровых фильтров.

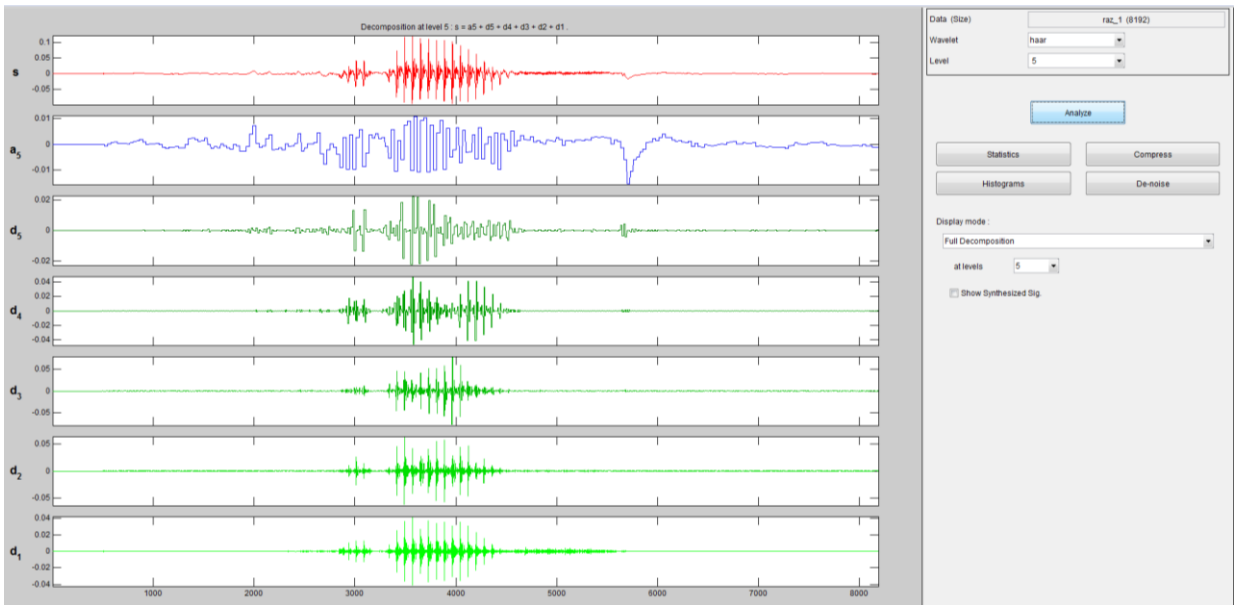

Рис. 1. Вейвлет-преобразование, Хаар, 5-й уровень

На рис. 1 показано вейвлет-преобразование сигнала, представляющего собой слово "Раз", с использованием вейвлета Хаара, или 1-го Добеши. Здесь:

s - исходный сигнал;

*a*5 - аппроксимирующий сигнал для 5-го уровня декомпозиции;

*d*1-*d*5 - детализирующие сигналы с 1-го по 5-й уровень декомпозиции.

Последующие уровни декомпозиции получаются путём применения вейвлет-преобразования к аппроксимирующему сигналу предыдущего уровня. Длина получившихся сигналов (аппроксимирующего и детализирующего) равна половине длины исходного сигнала.

Подобный подход называется кратномасштабным анализом (КМА).

На определённом уровне разложения сигнал представлен суммой аппроксимирующей составляющей и полученных на всех этапах, включая данный, детализирующих составляющих. Таким образом, на картинке $s=a5+d5+d4+d3+d2+1$.

Нулевые моменты. Изменения сигнала хорошо локализуются НВП, где используются вейвлеты, в которых достаточно нулевых моментов. При значениях времени, соответствующих гладким частям сигнала, значения НВП малы, а при изменениях сигнала относительно велики.

Например, при анализе сигнала, состоящего из двух прямых, значения НВП с использованием вейвлета «Мексиканская шляпа» (два нулевых момента) лучше локализованы вокруг точки пересечения этих двух прямых, чем при использовании вейвлета Хаара(один нулевой момент).

Поскольку звуковой сигнал обладает сложной структурой, выбор вейвлета для анализа не имеет однозначных критериев. В данной работе использован вейвлет Хаара (или 1-й Добеши).

При КМА сигнал последовательно раскладывается на аппроксимирующую и детализи-

рующую составляющие, длина которых равна половине длины исходного сигнала. На первом шаге операция производится с самим сигналом, на последующих – с его аппроксимацией. При использовании пакетных вейвлетов детализирующий сигнал также подвергается декомпозиции. Результат преобразования представляет собой все аппроксимирующие и детализирующие сигналы, полученные на определённом уровне декомпозиции. Длина такой последовательности равна длине исходного сигнала.

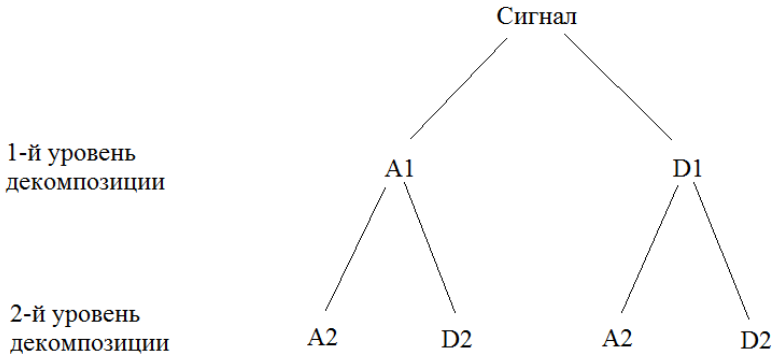

Рис. 2. Дерево декомпозиции:
A1 - аппроксимирующий сигнал, полученный на первом уровне декомпозиции;
D1 - аппроксимирующий сигнал, полученный на первом уровне декомпозиции;
A2 - аппроксимирующий сигнал, полученный на втором уровне декомпозиции;
D2 - аппроксимирующий сигнал, полученный на втором уровне декомпозиции

Далее приведён результат пакетного вейвлет-преобразования слова «Раз» (рис. 3) с использованием пакетного вейвлета Хаара для 2-го уровня декомпозиции.

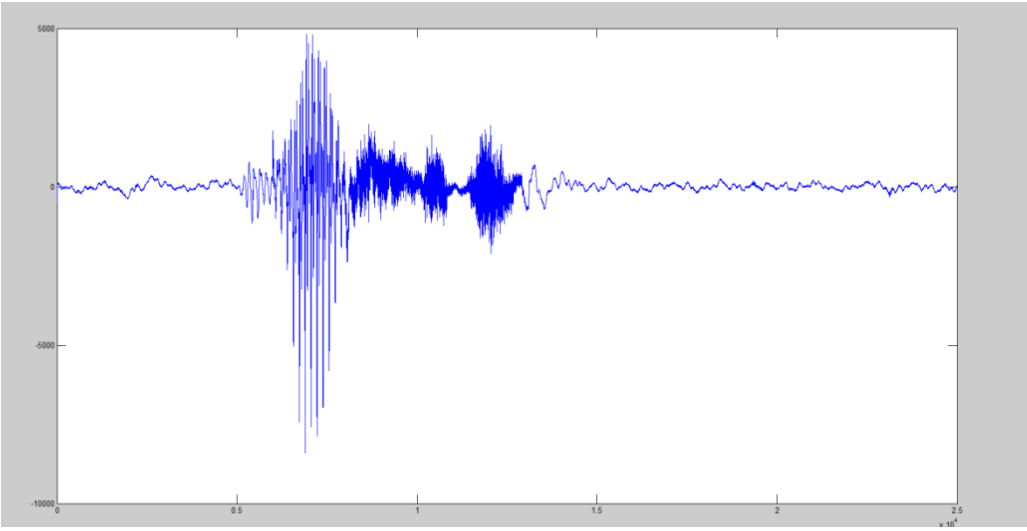

Рис. 3. Исходный сигнал "Раз"

Результирующий сигнал, таким образом, состоит из четырёх частей – аппроксимации *A*2 и детализации *D*2, полученных из аппроксимации 1-го уровня декомпозиции *A*1, и того же для детализирующего сигнала 1-го уровня *D*1. Эти части записаны одна за другой в одну последовательность, занимая каждая одну четверть сигнала, и при необходимости могут анализироваться как совместно, так и раздельно.

Распознавание речи требует сравнения сигналов между собой. В связи с тем, что даже одинаковые слова у разных людей могут различаться по времени произношения, находят применение различные алгоритмы сравнения временных последовательностей. В данной ра-

боте используется алгоритм динамической трансформации временной шкалы, или Dynamic Time Warping (**DTW**). Данный алгоритм позволяет находить степень похожести различных по времени и смещённых один по отношению к другому сигналов.

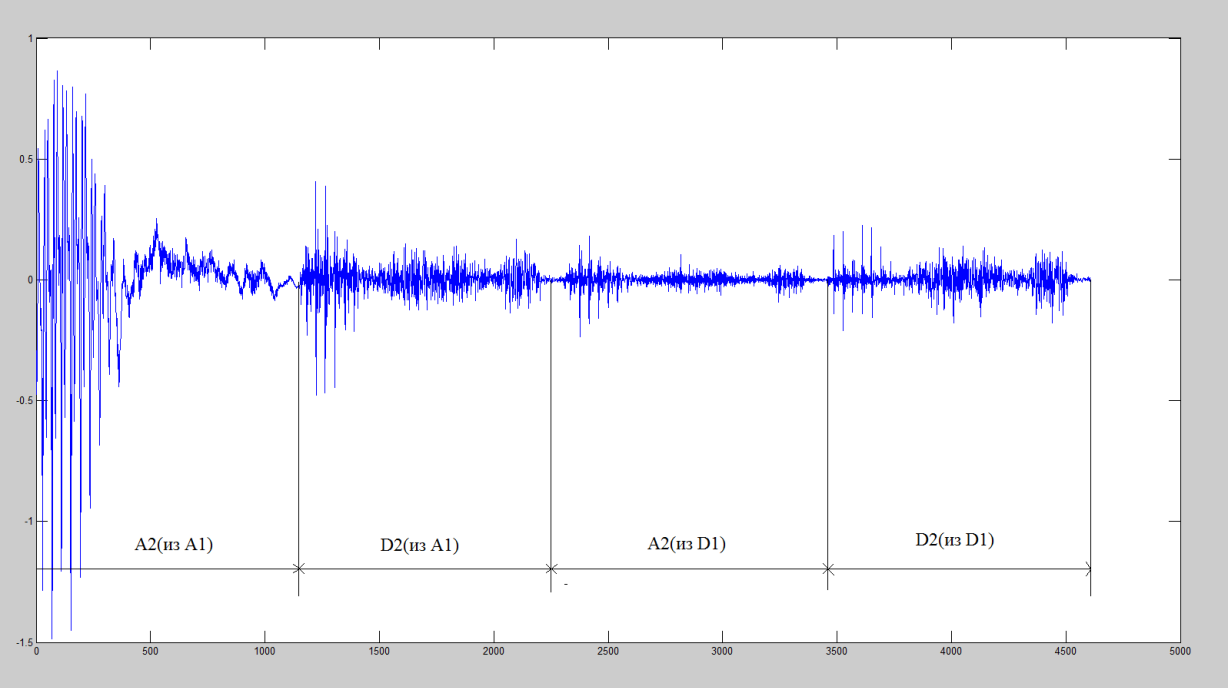

Рис. 4. Пакетное ДВП

Классический алгоритм DTW заключается в следующем:

Пусть даны две последовательности Q и C (временных ряда) длиной n и m соответственно:

$$Q = q_1,\ q_2, \dots, q_n\ ,\ C = c_1,\ c_2, \dots, c_m\ .$$

Классический DTW алгоритм по этим последовательностям строит путь наименьшей стоимости. Поясним, что это значит.

Определим матрицу $\Omega^{n\times m}$ так, чтобы её элемент (i, j) соответствовал расстоянию между i-м и j-м элементами последовательностей Q и C, т.е. соответствовал выравниванию между q_i и c_j. Мы будем брать евклидово расстояние:

$$d(q_i, c_j) = (q_i - c_j)^2.$$

По матрице построим некоторый путь W. Этот путь выражает соответствие между Q и C, k-й элемент W определяется как $w_k = (i, j)$. Далее под $d(w_k)$, где $w_k = (i,j)_k$, будем понимать $\mathrm{d}(q_i, c_j)$, т. е.

$$d(w_k) = \mathrm{d}(q_i, c_j) = (q_i - c_j)^2.$$

Итак, мы имеем $W = w_1, w_2, \dots, w_k, \dots, w_K$, где K - длина пути,

K удовлетворяет следующему условию:

$$\min(m,n) \le K < m + n - 1.$$

Пусть путь W удовлетворяет следующим условиям:

– Граничные условия

Обычно предполагают, что $w_1 = (1, 1)$ и $w_K = (n, m)$, т. е. начало и конец W находятся на диагонали в противоположных углах Ω;

– Непрерывность

Пусть $w_k = (a, b)$ и $w_k{-}1 = (p, q)$. Тогда

$$a - p \le 1, b - q \le 1.$$

Это ограничение нужно, чтобы в шаге пути W участвовали только соседние элементы матрицы (включая соседние по диагонали);

– Монотонность

Пусть $w_k = (a, b)$ и $w_k{-}1 = (p, q)$. Тогда

$$a - p \geq 1, b - q \geq 1.$$

Это ограничение нужно, чтобы точки W монотонно перемещались во времени. Путей, удовлетворяющих этим трем условиям, может быть очень много. Однако нам нужен путь, на котором достигается минимум стоимости пути:

$$\mathrm{DTW}(Q,C) = \min\left\{\frac{1}{K}\sqrt{\sum_{k=1}^{K} d(w_k)}\right\}.$$

Знаменатель K нужен для того, чтобы учесть различную длину W.

Таким образом, путь наименьшей стоимости (выравнивающий путь) для последовательностей Q и C это путь W, на котором достигается минимум стоимости пути DTW(Q,C).

Классичесский DTW алгоритм поиска пути минимальной стоимости рекурсивно находит длину пути наименьшей стоимости i, j до каждого элемента матрицы:

$$\gamma_{\mathrm{i,j}} = \mathrm{d}(w_{\mathrm{i,j}}) + \min(\gamma_{\mathrm{i,j-1}}, \gamma_{\mathrm{i-1,j}}, \gamma_{\mathrm{i-1,j-1}}).$$

Таким образом, DTW позволяет оценить степень похожести сигналов.

По теореме Котельникова, аналоговый сигнал может быть восстановлен однозначно и без потерь по своим дискретным отсчетам, взятым с частотой строго большей удвоенной верхней частоты. Человеческому голосу соответствует диапазон частот 300–4000 Гц, следовательно, для того чтобы восстановить голосовой сигнал без потерь, необходимо использовать частоту дискретизации, большую 8 кГц. Для проведения экспериментов использовалась частота 11025 Гц.

Скрытой марковской моделью (**СММ**) называется модель, обладающая следующими характеристиками: модель имеет N состояний, в каждом из которых она принимает одно из M значений. Вероятности переходов между состояниями определяются матрицей A=$\{a_{ij}\}$, где a_{ij} - вероятность перехода из i в j состояние. Вероятности выпадения определённых значений в каждом состоянии определяются матрицей B=$\{b_j(k)\}$, где $b_j(k)$ - вероятность выпадения k-го значения в j-м состоянии. Также модель содержит матрицу вероятностей $\pi = \{\pi_i\}$, где π_i - вероятность, что в начальный момент система в i-м состоянии.

Таким образом, скрытая марковская модель может быть описана как $\lambda = \{A, B, \pi\}$.

Для СММ решаются 3 задачи.

Первая задача – вычисление вероятности появления наблюдаемой последовательности для определённой СММ. Для решения первой задачи применяют алгоритм прямого хода и алгоритм обратного хода.

Вторая задача - выбор последовательности состояний определённой СММ, с наибольшей вероятностью порождающей указанную последовательность наблюдений. Для решения второй задачи применяют алгоритм Витерби.

Третья задача – подбор параметров СММ так, чтобы максимизировать вероятность появления определённой последовательности наблюдений. Для решения третьей задачи применяют алгоритм Баума-Уэлча.

При распознавании изолированных слов скрытые марковские модели могут соответствовать распознаваемым словам, при этом состояния модели соответствуют участкам сигнала. Значения, которые СММ принимает в этих состояниях, могут соответствовать определённым параметрам сигнала на заданном участке.

В целях оценки эффективности работы системы распознавания изолированных слов, использующей пакетные вейвлеты и скрытые марковские модели, был разработан следующий алгоритм:

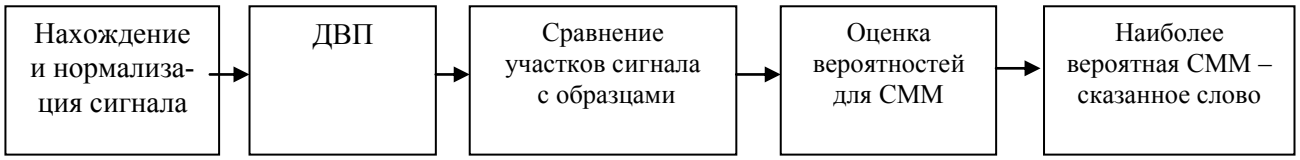

Рис. 5. Схема работы системы

Реализован алгоритм поиска начала сигнала по определённому уровню. Началом считается точка, отстоящая на некоторой дистанции от того момента, когда сигнал достигает заданного уровня. Уровень и отступ определены экспериментально.

Нормализация сигнала заключается в делении сигнала на максимальное его значение. Это позволяет работать с сигналами различной громкости, поскольку полученные значения будут между -1 и 1.

Для сравнения сигналов использован алгоритм Dynamic time warping, который сравнивает две числовые последовательности и в качестве результата выдаёт число, характеризующее их похожесть. Число тем меньше, чем более похожи последовательности.

Таким образом, значение сказанного слова определяется, исходя из того, какой из ранее записанных образцов наиболее похож на него.

Система настроена для распознавания слов «Раз», «Два», «Три», с использованием 9 образцов (по 3 для каждого слова). Также был опробован вариант с распознаванием десяти цифр.

Для распознаваемых слов построены СММ со следующими параметрами:

4 состояния, в каждом по 9 значений.

Состояния – 4 участка сигнала, соответствующие аппроксимирующим и детализирующим последовательностям.

Вероятности переходов выставлены таким образом, что происходит последовательный переход между состояниями.

На каждом из участков определяется похожесть на один из образцов с помощью алгоритма Dynamic Time Warping. Образцы – соответствующие отрезки сигналов – образцов. Таким образом, наблюдаемые последовательности, с которыми работают СММ, представляют собой четыре значения, соответствующие сигналам-образцам, на которые соответствующие участки наиболее похожи.

После получения такой последовательности решается первая задача СММ, т.е. определяется, какая из СММ с наибольшей вероятностью её генерирует. Слово, которое соответствует данной СММ, и является результатом распознавания.

Данный алгоритм реализован на языке Java, для СММ использована библиотека с открытым исходным кодом Jahmm v0.6.1.

Применение СММ обеспечило следующие преимущества:

- возможность использования нескольких образцов одного и того же слова. СММ возможно настроить таким образом, что вероятности появления сигналов, соответствующих одному слову, равны, и при этом больше, чем вероятности для других слов.
- распознаваемый сигнал может быть похож на различные образцы на разных участках, СММ позволяет учитывать только то, к какому слову относятся образцы. Это влияет на результат в случае, когда наблюдаются совпадения с различными образцами, и важно то, к какому слову каждый образец относится.

С целью оценки эффективности распознавания для различных конфигураций системы создана база образцов, состоящая из десяти записей для трёх слов "Раз", "Два", "Три".

Первая строка таблицы показывает количество и процент корректно распознанных слов

при использовании пакетного вейвлет-преобразования второго уровня декомпозиции, т.е. сравниваются последовательности, содержащие и аппроксимирующие, и детализирующие сигналы. Кроме того, деление сигнала на четыре участка позволяет независимо сравнивать эти компоненты. Данный подход показал наилучшие результаты.

Таблица 1

Вейвлет-преобразование	Процент распознавания	Количество распознанных слов
Пакетный Хаар, 2 уровень	86%	(26/30)
A2+D2+D1	70%	(21/30)
A4+D4	56%	(17/30)
A2+D2	60%	(18/30)
A4	66%	(20/30)
D4	33%	(10/30)
A3	46%	(14/30)
D3	43%	(13/30)
A2	50%	(15/30)
D2	53%	(16/30)
A1	46%	(14/30)
D1	56%	(17/30)
Исходный сигнал	40%	(12/30)

Далее показаны результаты распознавания для аппроксимирующих (Ai) и детализирующих (Di) сигналов различных уровней декомпозиции, а также их сочетаний. Наиболее близкий к пакетному вейвлету результат получен для сочетания аппроксимирующих и детализирующих компонент вейвлет-преобразования, также на втором уровне декомпозиции.

Выводы

С целью оценки эффективности распознавания при большем количестве распознаваемых слов записана база из 100 записей (слова от "Раз" до "Десять"), по 10 записей для каждого слова. Количество образцов в системе для распознавания каждого слова по-прежнему три (всего тридцать образцов). Каждому слову соответствует своя скрытая марковская модель, таким образом, всего десять моделей. Используется пакетный вейвлет Хаара, второй уровень декомпозиции.

Таблица 2

Распознавание	Процент распознавания	Количество распознанных слов
10 слов	65%	(65/100)
Раз, два, три	83%	(25/30)
Трёхбуквенные слова	84%	(42/50)

Результирующая таблица показывает, что процент распознанных слов существенно ниже. Тем не менее, для слов "Раз", "Два", "Три" процент уменьшился незначительно. Кроме того, тот же результат наблюдается для прочих коротких слов, что позволяет предположить, что характеристики полученной системы распознавания оптимальны для работы с подобными сигналами.

В данной работе, посвящённой распознаванию речи, мы опробовали алгоритм, сочетающий вейвлет-преобразование в различных конфигурациях, алгоритм динамической трансформации временной шкалы и скрытые марковские модели. Алгоритм применён для распознавания изолированных слов, произносимых одним диктором, при различных объёмах словаря системы и с использованием разных результатов вейвлет-преобразования.

Библиографический список

1. **Яковлев, А.Н.** Я 474 Введение в вейвлет-преобразования: учеб. пособие. – Новосибирск: Изд-во НГТУ, 2003. – 104 с.
2. **Сергиенко, А. Б.** Цифровая обработка сигналов / А. Б. Сергиенко. – СПб.: Питер,2002. –602 с.
3. **Штарк, Г.Г.** Применение вейвлетов для ЦОС / Г.Г. Штарк. – М.: Техносфера, 2007. – 192 с.
4. **Смоленцев, Н. К.** Основы теории вейвлетов. Вейвлеты в Matlab / Н. К. Смоленцев. – М.: ДМК Пресс, 2005. – 304 с.
5. Texas Instruments C5505 Teaching Materials.
6. Непрерывное wavelet преобразование [Электронный ресурс] // Habrahabr. URL: http://habrahabr.ru/post/103899/
7. Динамическое программирование в алгоритмах распознавания речи [Электронный ресурс]// Habrahabr. URL: https://habrahabr.ru/post/135087/
8. Применение скрытых марковских моделей для распознавания речи [Электронный ресурс]// Компьютерное распознавание и порождение речи. URL: http://speech-text.narod.ru/chap4.html

Дата поступления
в редакцию 08.02.2016

O. Korelin, E. Leonova, P. Tanonov

WAVELET TRANSFORM AND HIDDEN MARKOV MODELS IN SPEECH RECOGNITION

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

Purpose: Create a speech recognition algorithm using a combination of methods.
Design/methodology/approach: The article deals with wavelet transform and hidden Markov models in speech recognition. Considered various options for system configuration using specific results of the wavelet transform and different volume of the dictionary. Dynamic time warping algorithm is used for signal comparison.

Key words: speech recognition, wavelet transform, dynamic time warping, hidden Markov models.