УДК 681.142.66

Т.В. Ермоленко
Донецкий государственный институт искусственного интеллекта,
etv@iai.donetsk.ua, Украина

Использование непрерывного вейвлет-преобразования при распознавании вокализованных участков речевого сигнала

В статье рассматривается вопрос о способе представления речевого сигнала через его вейвлет-разложение, обладающее большей информативностью по сравнению со спектральным представлением. Предложена методика построения системы признаков распознавания фонем, соответствующих вокализованным звукам, на основе непрерывного вейвлет-преобразования с использованием DOG-вейвлета в качестве материнского. Помимо этого приведены алгоритмы для получения коэффициентов непрерывного вейвлет-спектра и выделения квазипериода сигнала.

Актуальность работы

Построение мобильных распознавателей речи является актуальной задачей современной техники. Будучи встроенными в мобильные устройства, такие распознаватели способны существенно облегчить взаимодействие пользователя и системы.

Для построения распознающей системы, обладающей устойчивостью к возможным вариациям свойств диктора и окружающей обстановки, актуальной является задача преобразования входного речевого сигнала в набор акустических параметров и формирования из них компонент вектора характерных признаков, которые в дальнейшем будут использованы для распознавания.

Анализ исследований и публикаций

Перед разработчиками встает ряд проблем, среди которых необходимо отметить нелинейное растяжение произносимого во времени и значительную нестабильность речевого сигнала. В настоящее время для преодоления этой проблемы используются методы динамического искривления во времени (DTW), позволяющие определять расстояние между различными реализациями речевого сигнала с учетом возможной повторяемости элементов этих реализаций [1]. Этот подход требует значительного объема вычислений, поскольку предполагает сравнение распознаваемого образа со всеми элементами словаря эталонов.

Поэтому гораздо удобнее оперировать ограниченным набором элементов речи, описывающим весь лексический словарь, в частности, на уровне фонем.

Большинство методов акустико-фонетического анализа используют спектральное либо амплитудно-формантное представление речевого сигнала, которое преобразуют в набор признаков для описания динамики сигнала [2]. Представление сигнала с использованием вейвлетов как средства многомасштабного анализа позволяет выделять одновременно как основные характеристики сигнала, так и короткоживущие высокочастотные явления. Это свойство является существенным преимуществом в задачах обработки речевого сигнала по сравнению с оконным преобразованием Фурье, где, варьируя ширину окна, приходится выбирать масштаб явлений, которые необходимо выделить в сигнале.

С точки зрения распознавания речи вейвлет-преобразование можно рассматривать как некоторый фильтр. Используя различный масштаб, можно получить речевой сигнал, отфильтрованный в различных частотных диапазонах. В работе [3], а также работах Ф.Г. Бойкова, Т.К. Старожиловой, посвященных обработке речи с применением вейвлет-анализа, используется масштаб $s = 2^k$, что дает большую ширину частотных диапазонов и сильное огрубление сигнала на соответствующих уровнях. Возникает проблема выбора масштаба с целью получения более гибкой частотной решетки. Помимо этого, для построения признаков распознавания сигнал нарезают на перекрывающиеся окна постоянной длины. Но с учетом неинвариантности к сдвигу вейвлет-преобразования значения вейвлет-коэффициентов на разных участках одной и той же фонемы колеблются, что приводит к ухудшению точности распознавания.

Цель и задачи исследования

Целью данной работы является построение на основе непрерывного вейвлет-преобразования набора признаков распознавания вокализованных звуков, инвариантных к сдвигу внутри фонемы и устойчивых к изменению абсолютного уровня входного сигнала и уровня записи.

Задачи исследования: разработать алгоритм получения представления речевого сигнала через его вейвлет-разложение с возможностью настройки масштабного коэффициента под диктора; разработать метод выделения квазипериода по вейвлет-спектру; сформировать набор признаков распознавания.

Постановка задачи

Получить представление исходного сигнала в пространстве (временной масштаб, временная локализация) с возможностью настройки под диктора;

выделить период основного тона;

на основе полученного представления построить систему признаков, удовлетворяющих следующим требованиям:

1. Признаки должны быть инвариантны к сдвигу.
2. Изменение абсолютного уровня входного сигнала и изменение уровня записи не должно заметно влиять на получаемые численные значения признаков.
3. Система признаков должна допускать представление в виде, пригодном для дальнейшей обработки PC, т.е. представлять собой вектор числовых значений фиксированной длины.

Ниже приведено подробное описание последовательности преобразований, позволяющих получить наборы признаков из речевого сигнала.

Представление сигнала через его вейвлет-разложение

Если конструировать базис функционального пространства $L^2(R)$ с помощью непрерывных масштабных преобразований и переносов вейвлета $\psi(t)$ с произвольными значениями базисных параметров – масштабного коэффициента *a* и параметра сдвига *b* [4-6]:

$$\psi_{ab}(t) = |a|^{-1/2}\psi\left(\frac{t-b}{a}\right), \qquad a,b \in R, \quad \psi \in L^2(R),$$

то на его основе можно записать непрерывное вейвлет-преобразование функции $f(t)$:

$$CWT(a,b) = |a|^{-1/2}\int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right)dt = \int_{-\infty}^{\infty} f(t)\psi_{ab}(t)dt$$

Результатом вейвлет-преобразования сигнала является двумерный массив значений коэффициентов *CWT(a,b)*. Распределение этих значений в пространстве *(a,b)* (временной масштаб, временная локализация) дает информацию об эволюции относительного вклада компонент разного масштаба во времени и называется вейвлет-спектром [7].

В своих исследованиях мы использовали хорошо известный DOG-вейвлет, построенный на основе гауссовской функции (Difference of Gaussians) [6]:

$$\psi(t) = \exp\left(-\frac{|t|^2}{2}\right) - 0{,}5\exp\left(-\frac{|t|^2}{8}\right).$$

Отображение сигнала на плоскости *(a;b)*, полученное посредством непрерывного преобразования, является избыточным: во-первых, сигналы физического происхождения имеют конечную длину реализации, а, стало быть, ограниченную полосу частот, а во-вторых, результат преобразования (функция *CWT(a,b)*) при определенных условиях решаемой задачи допускает наличие погрешностей анализа. Кроме того, процедура непрерывного преобразования, теоретически исполняемая для $t \in R$, $b \in R$, $a \in R^+$, очевидно, является нереализуемой на практике, в связи с чем возникает необходимость в использовании процедур квантования переменных, и, таким образом, перехода к процедуре представления сигналов с помощью рядов [4].

Ниже предложен следующий алгоритм непрерывного вейвлет-преобразования.

Известны временные отсчеты сигнала *f(n)* в моменты времени $n \in [0,N-1]$, где N – количество измерений. Расстояние между отсчетами dt – величина, обратная частоте дискретизации. В наших исследованиях частота дискретизации сигнала составляла 22050 Гц. Алгоритм позволяет выбирать три параметра, определяющие масштаб и глубину разложения сигнала: a_0, j_0, j_m, что дает возможность варьировать набор задействованных масштабов и сосредотачиваться на тех из них, где наблюдаются особенности (нерегулярность) сигнала.

Зная основную формулу для материнского вейвлета, необходимо построить базисные функции, так называемые «масштабированные» вейвлеты, имеющие вид:

$$\psi_{jn}(n') = s_j^{-1/2}\psi\left(\frac{(n'-n)dt}{s_j}\right),$$

где $s_j = a_0^{-j}$ – параметр, обратный частоте и определяющий масштаб, $a_0 > 1$.

Как было сказано выше, вейвлет-преобразование является сверткой сигнала *f(n)* с функцией-вейвлетом. Значения вейвлет-коэффициентов вычисляются в двойном цикле. Внешний – по j от j_0 до j_m, внутренний – по n от 0 до $N-1$, в котором вычисляется сумма:

$$W_{jn} = \sum_{n'=0}^{N-1} f(n')\psi_{jn}(n')dt\ .$$

В своих исследованиях мы использовали следующие значения параметров: $a_0 = 4$ либо $a_0 = 3{,}5$ (для женского либо мужского голоса соответственно); $j_0 = 5$; $j_m = 11$.

Выделение квазипериода

Признаки для распознавания вокализованных звуков будем строить на каждом участке, соответствующем одному квазипериоду, в связи с чем возникает необходимость его выделения. Это довольно легко сделать, используя вейвлет-коэффициенты на масштабе, соответствующем $j_0 = 5$. Дело в том, что вейвлет-коэффициенты на этом уровне разложения представляют собой сильно сглаженную и усредненную относительно сигнала функцию.

Введем величину, являющуюся аналогом первой производной для W_{5n}, $\Delta W_{5n} = W_{5n+1} - W_{5n}$. Будем считать началами квазипериодов отсчеты исходного сигнала с такими номерами n, при которых ΔW_{5n} меняет знак с «+» на «–».

Построение вектора признаков

Как известно [7], [8], с точки зрения распознавания речи, вейвлет-преобразование можно рассматривать как некоторый фильтр. Используя различный масштаб $s_j = a_0^{-j}$, можно получить речевой сигнал, отфильтрованный в различных частотных диапазонах.

Используем этот факт для построения характерных признаков для вокализованных звуков. Далее, на каждом уровне j найдем максимум по модулю W_{jn} и делим на него все коэффициенты разложения на этом уровне. Это позволит повысить устойчивость системы признаков к изменению абсолютного уровня входного сигнала и уровня записи. Обозначим полученные величины $\tilde{W}_{jn}$. Далее строим разности $D_{jn} = \tilde{W}_{j+1n} - \tilde{W}_{jn}$. Затем формируем вектор признаков с компонентами P_j:

$$P_j = \frac{1}{k_{i+1} - k_i} \sum_{n=k_i}^{k_{i+1}} \left| D_{jn} \right| \qquad (1)$$

где k_i и k_{i+1} – индексы отсчетов, соответствующих началу и концу i-го квазипериода, j меняется от j_0 до j_m. Выражение (1) определяет вклад каждой частотной полосы, соответствующей уровню j, в сигнал.

Локальные максимумы вейвлет-коэффициентов на каждом рассматриваемом уровне несут информацию о форме сигнала и являются информативными. В построенный вектор признаков добавляем еще $j_m – j_0 – 1$ компонент, являющихся количеством локальных максимумов вейвлет-коэффициентов, посчитанным на каждом уровне j (j от j_0+1 до j_m).

Спектры звуков, несмотря на их вариативность в разных реализациях и на разных спектральных срезах одной фонемы, имеют большое сходство для одной фонемы и значительно отличаются для разных фонем. Это отличие имеет стабильный характер и сохраняется от реализации к реализации, значит, по вектору признаков, построенному таким образом, можно будет адекватно провести классификацию.

Обсуждение полученных результатов

В эксперименте участвовало 7 дикторов. Полученные признаки позволяют безошибочно выделять 4 широких класса фонем: гласные (“а”,“и”,“о”,“у”,“э”), сонорные (“л”,“м”,“н”), смычные взрывные звонкие (“б”,“в”,“г”,“д”) и шумные звонкие согласные (“ж”, “з”) звуки. Внутри класса гласных точно определяются “а”,“и”, “э”. Внутри класса сонорных “л” распознается с вероятностью 80 %, точность распознавания “б”,“в”,“г”,“д” составила 75 %.

Выводы

1. Предложенная система признаков, построенная с помощью нормированных величин (1), устойчива к изменению абсолютного уровня входного сигнала и уровня записи.

2. За счет того, что анализируемый сигнал разбит не на окна постоянной длины, а на сегменты, соответствующие началу и концу квазипериода для конкретного диктора, полученный набор признаков распознавания инвариантен к сдвигу.
3. Система признаков представляет собой вектор числовых значений фиксированной длины и пригодна для построения кодовой книги, обучения скрытых моделей Маркова и дальнейшего распознавания речи.

В данной работе новыми являются следующие положения:

– предложен алгоритм непрерывного вейвлет-преобразования, позволяющий настраивать масштаб разложения и соответственно ширину частотных диапазонов под конкретного диктора;
– предложен достаточно простой и эффективный алгоритм выделения периода основного тона на вокализованных участках сигнала;
– представлен способ построения вектора признаков на основе разностей (1), характеризующих вклад различных частотных диапазонов в анализируемый сигнал, устойчивый к амплитудным изменениям входного сигнала и инвариантный к сдвигу.

Литература

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наук. думка, 1987. – 262 с.
2. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов. – М.: Радио и связь, 1981. – 496 с.
3. Юрков П.Ю., Федоров В.М., Бабенко Л.К. Распознавание гласных фонем с помощью нейронных сетей. – Красноярск: Тезисы доклада 7 Всероссийского семинара «Нейроинформатика и ее приложения», 1999. – 159 с.
4. Добеши И. Десять лекций по вейвлетам. – Москва; Ижевск: НИЦ «регулярная и хаотическая динамика», 2004. – 464 с.
5. Астафьева Н.М. Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук. – 1996. – Т. 166, № 11. – С. 1145-1170.
6. Дремин И.М., Иванов О.В., Нечитайло В.А. Вейвлеты и их использование // Успехи физических наук. – 2002. – Т. 171, № 5. – С. 465-500.
7. Воробьев В.И., Грибунин В.Г. Теория и практика вейвлет-преобразования. – СПб.: Изд-во ВУС, 1999. – 208 с.
8. Новиков Л.В. Спектральный анализ сигналов в базисе вейвлетов // Научное приборостроение. – 2000. – Т. 10, № 3. – С. 57-64.
9. Златоустова Л.В. Фонетические единицы русской речи. – М.: Московский университет, 1981. – 108 с.
10. Новиков Л.В. Основы вейвлет-анализа сигналов. Учебное пособие. – СПб.: Изд-во ООО «МОДУС+», 1999. – 152 с.

Т. В. Єрмоленко

Використання безперервного вейвлет-перетворення розпізнаванні вокалізованих участків мовленнєвого сигналу

У статті розглядається питання про спосіб представлення мовленнєвого сигналу через його вейвлет-розкладання, що володіє більшою інформативністю в порівнянні зі спектральним представленням. Запропоновано методику побудови системи ознак розпізнавання фонем, що відповідають вокалізованим звукам, на основі безперервного вейвлет-перетворення з використанням DOG-вейвлету як материнського. Крім цього приведені алгоритми для одержання коефіцієнтів безперервного вейвлет-спектру і виділення квазиперіоду сигналу.

Tatyana V. Yermolenko

The continuous wavelet-transformation using in the tasks of a vocal sounds recognition in a speech signal

In a paper the speech signal representation through its wavelet-decomposition is considered. The technique of a features system construction for a vocal sounds recognition is proposed. This method uses a continuous wavelet-transformation with DOG- wavelet ered as mother. Moreover the algorithms for deriving coefficients of a continuous wavelet-spectrum and a signal quasiperiod computation are produced.

Статья поступила в редакцию 20.07.04.