Применение вейвлет-анализа сигнала в системе распознавания речи

Бойков Ф.Г.
Старожилова Т.К.

Введение

В данной работе описана серия экспериментов по применению вейвлет анализа для задач распознавания речи. Получены оценки точности распознавания изолированных слов и слитно произносимых последовательностей цифр при использовании блока предобработки, основанного на вейвлет-преобразовании речевого сигнала.

Применение вейвлетов в задачах обработки и распознавания речи продиктовано особенностями речевого акустического сигнала. Вейвлеты, как средство многомасштабного анализа позволяют выделять, одновременно как основные характеристики сигнала, так и короткоживущие высокочастотные явления в речевом сигнале. 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.

Получение дополнительной информации с разных масштабов времени и разных масштабов разрешения сигнала может улучшить точность распознавания речи. Кроме того, считается [1], что человеческое ухо устроено так, что при обработке звукового сигнала, оно передает мозгу вейвлет-образ сигнала. Колебания амплитуды давления передаются от барабанных перепонок не мембрану и далее распространяются по всей длине завитка внутреннего уха. Завиток скручен в виде спирали во внутреннем ухе. Если представить, что завиток распрямлён в некоторый сегмент, а вместе с ним и распрямлена мембрана, то можно показать, что результирующее преобразование сигнала будет с точностью до константы совпадать с вейвлет-преобразованием.

Применение вейвлет-анализа в задачах обработки и распознавания речи

Вейвлет-анализ – это исследование сигнала $f(t)$ при помощи разложения по системе базисных функций. Сигнал $f(t)$ интерпретируется, как функция из $\mathbf{L}^2(\mathbf{R})$, а в качестве базиса используется система функций $\psi_{a,b}(t) = \psi\left(\frac{b-t}{a}\right)$, занумерованных не целыми числами, а двумя непрерывными параметрами. Эта система получается из фиксированной функции $\psi(t)$ всевозможными сдвигами и растяжениями. Функция $\psi(t)$ называется вейвлетом (по-английски – wavelet; в русской математической литературе используется также термин «всплеск»), если:

1) $\psi(t)$ непрерывна;
2) $\psi(t)$ интегрируема на всей прямой;

3) $\int_{-\infty}^{\infty} \psi(t)\, dt = 0$

Вейвлет-преобразованием $f(t)$ называется функция двух переменных

$$Wf(b,a) = \frac{1}{a}\int_{-\infty}^{\infty} f(t)\psi\left(\frac{b-t}{a}\right) dt$$

В отличие от традиционного преобразования Фурье, вейвлет-преобразование определено неоднозначно: каждому вейвлету соответствует свое преобразование. Условие 3 означает, что Фурье-образ $\hat{\psi}(\omega)$ вейвлета обращается в 0 при $\omega = 0$; это нужно для того, чтобы в Фурье-области вейвлет был локализован вокруг некоторой ненулевой частоты ω_0. В качестве анализирующих вейвлетов обычно выбираются функции, хорошо локализованные также и в «пространственной области» (т.е. по t).

t

w

Центры частотно-временной локализации для вейвлетов на плоскости спектрограммы показаны на рис.1.

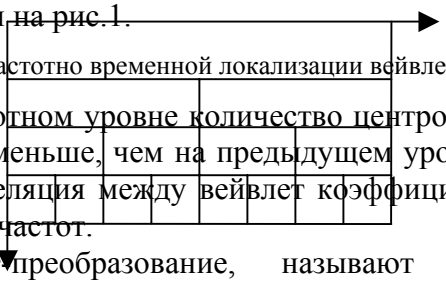

Рисунок 1. Центры частотно временной локализации вейвлет коэффициентов.

На каждом частотном уровне количество центров частотно-временной локализации в два раза меньше, чем на предыдущем уровне, частота которого выше. Существует корреляция между вейвлет коэффициентами как по шкале времени, так и по шкале частот.

Такое вейвлет преобразование, называют непрерывным. Его выполнение требует больших вычислительных затрат. Желательно было бы иметь разложение с вейвлетами в качестве ортогональных базисных функций. Такие вейвлеты существуют, и очень полезны в задачах сжатия информации и подавления шумов. Схема построения таких вейвлетов связана с исчерпанием пространства сигналов системой вложенных подпространств, отличающихся друг от друга только перемасштабированием независимой переменной. Такая система называется многомасштабным анализом (multiresolution analysis).

В качестве пространства сигналов будем рассматривать $\mathbf{L^2(R)}$ – пространство комплекснозначных функций $f(t)$ на прямой, для которых $\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty$. В этом пространстве определено скалярное произведение функций по формуле $< f, g > = \int_{-\infty}^{\infty} f(t)\bar{g}(t)dt$. Число $\|f\| = \sqrt{< \mathrm{f}, \mathrm{f} >}$ называется нормой функции $f(t)$. Базисом в пространстве $\mathbf{V} \subset \mathbf{L^2(R)}$ называется такая система функций $\{\mathbf{v}_i(t)\}$, что любая функция

$\mathbf{v}(t) \in \mathbf{V}$ единственным образом записывается в виде $\mathbf{v}(t) = \sum \mathrm{c_i} \mathbf{v}_i(t)$. Базис называется ортонормированным, если $< \mathbf{v}_i(t), \mathbf{v}_j(t) > = \delta_{ij}$. В этом случае $\mathbf{c}_i = < \mathbf{v}(t), \mathbf{v}_i(t) >$.

Популярный пример ортобазиса в пространстве периодических функций – базис Фурье $\mathbf{v}_n(t) = e^{-i \cdot n \cdot t}$. В пространстве $\mathbf{L^2(R)}$ также есть много классических базисов – Эрмита, Лаггера и др.

Ортогональным многомасштабным анализом в пространстве $\mathbf{L^2(R)}$ называется система подпространств $\mathbf{V}_j \subset \mathbf{L^2(R)}, j = 0, \pm 1, ...$, удовлетворяющая следующим условиям.

1) $\mathbf{V}_j \subset \mathbf{V}_{j+1}$,
2) $v(t) \in \mathbf{V}_j \Leftrightarrow v(2t) \in \mathbf{V}_{j+1}$,
3) $v(t) \in \mathbf{V}_0 \Leftrightarrow v(t+1) \in \mathbf{V}_0$,
4) $\overline{\bigcup \mathbf{V}_j} = \mathbf{L^2(R)}, \bigcap \mathbf{V}_j = 0$,

если существует $\varphi(t) \in \mathbf{V}_0$ такая, что функции $\{\varphi(t-m)\}, \mathrm{m} = 0, \pm 1, ...$ образуют ортонормированный базис пространства $\mathbf{V}_0$.

Функция $\varphi(t)$ называется скейлинг-функцией (scaling function).

Пусть $\varphi(t) = 1$ на интервале $[0,1)$ и $\varphi(t) = 0$ вне этого интервала. Целочисленные сдвиги этой функции попарно ортогональны. $\mathbf{V}_0$ состоит из функций, постоянных на интервалах вида $[n, n+1)$, $\mathbf{V}_1$ – из функций, постоянных на интервалах вида $\left[\frac{n}{2}, \frac{n+1}{2}\right)$, $\mathbf{V}_{-1}$ – из функций, постоянных на интервалах вида $[2n, 2n+2)$, и т.д. Любую функцию из $\mathbf{L^2(R)}$ можно приблизить функциями такого вида. В современной терминологии $\varphi(t)$ называется скейлинг-функцией Хаара (Haar).

Будем считать, что пространство $\mathbf{V}_0$ состоит из сигналов, заданных «с разрешением 1». Тогда пространство $\mathbf{V}_j$ – сигналы, заданные с разрешением 2^{-j}. Любое $\mathbf{V}_j$ отличается от $\mathbf{V}_0$ только перемасштабированием. Поэтому пространство $\mathbf{V}_j$ порождено ортобазисом $\left\{2^{j/2} \varphi\left(2^j t - m\right)\right\}$. Например, $\mathbf{V}_{-1}$ порождено функциями вида $\left\{\frac{1}{\sqrt{2}} \varphi\left(\frac{t}{2} - m\right)\right\}$. Т.к. $\mathbf{V}_{-1} \subset \mathbf{V}_0$, функция $\frac{1}{\sqrt{2}} \varphi\left(\frac{t}{2}\right)$ обязана линейно выражаться через сдвиги $\varphi(t)$. Значит, существуют такие коэффициенты $\{h_k\}$, что

$$\varphi\left(\frac{t}{2}\right) = \sqrt{2} \sum\nolimits_k h_k \varphi(t-k).$$

Это – важнейшее в теории вейвлетов уравнение (уравнение масштабирования, уравнение рескейлинга, refinement equation, two-scale equation). Во многих практически важных случаях скейлинг-функция строится именно как решение этого функционального уравнения, в отличие от классических спецфункций, возникающих как решения дифференциальных уравнений. Неформально говоря, это связано с тем, что вейвлетный базис отслеживает изменение сигнала не только по времени, но и «по масштабу».

Для ортонормированных вейвлет-базисов существует быстрый алгоритм вычисления ортогонального вейвлет-преобразования.

Пусть $x(t) \in \mathbf{V}_0$, и нам даны коэффициенты x_n его разложения по сдвигам скейлинг-функции:

$$x(t) = \sum_n x_n \varphi(t-n)$$

Естественно считать версией масштаба 2 ортогональную проекцию $x(t)$ на подпространство $\mathbf{V}_{-1}$. Она задается набором скалярных произведений $x(t)$ с функциями из ортобазиса $\mathbf{V}_{-1}$, то есть величинами $c_r = \left\langle \mathrm{x(t)}, \frac{1}{\sqrt{2}}\varphi\left(\frac{t}{2}-r\right)\right\rangle$ и $c_r = \sum_{\mathrm{s}} \mathrm{h_s} x_{2r+s}$

Другими словами, проекция осуществляется путем свертки с фильтром h^* и прореживания вдвое. Заметим, что прореживание вдвое «встроено» в эту формулу (через индекс $2r+s$). Разумеется, это следствие выбора базиса в $\mathbf{V}_{-1}$.

В качестве деталей сигнала $x(t)$, исчезающих при переходе к масштабу 2, следует взять компоненту $x(t)$, ортогональную к сигналам масштаба 2, т.е. к пространству $\mathbf{V}_{-1}$. Мы видели, что имеет место разложение $\mathbf{V}_0 = \mathbf{V}_{-1} \oplus \mathbf{W}_{-1}$, где для любых функций $a(t) \in \mathbf{V}_{-1}, b(t) \in \mathbf{W}_{-1}$ выполнено $< a(t), b(t) > = 0$, и ортобазисом $\mathbf{W}_{-1}$ будет набор функций $\left\{\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}-m\right)\right\}$. Коэффициенты g_k имеют вид $g_k = (-1)^{\mathrm{k}} h_{1-k}$. Искомая проекция задается набором скалярных произведений $x(t)$ с функциями из ортобазиса $\mathbf{W}_{-1}$, то есть величинами $d_r = \left\langle \mathrm{x(t)}, \frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}-r\right)\right\rangle$ и $d_r = \sum_{\mathrm{s}} \mathrm{g_s} x_{2r+s}$, что равносильно свертке с фильтром g^* и прореживанию вдвое. Та же схема действует на любом масштабе. При любом j $\mathbf{V}_j = \mathbf{V}_{j-1} \oplus \mathbf{W}_{j-1}$, ортобазисом $\mathbf{W}_{j-1}$ будет $\left\{2^{(j-1)/2}\psi\left(2^{j-1}t-m\right)\right\}$. Совокупность же функций $\left\{2^{j/2}\psi\left(2^j t-m\right)\right\}$, где j и m пробегают все целые значения, будет базисом всего пространства $\mathbf{L^2(R)}$.

Если ввести матрицы H и G:

$$H = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & h_{-2} & h_{-1} & h_0 & h_1 & h_2 & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & h_{-2} & h_{-1} & h_0 & h_1 & h_2 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & h_{-2} & h_{-1} & h_0 & h_1 & h_2 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

$$G = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & g_{-2} & g_{-1} & g_0 & g_1 & g_2 & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & g_{-2} & g_{-1} & g_0 & g_1 & g_2 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & g_{-2} & g_{-1} & g_0 & g_1 & g_2 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

Тогда условия ортогональности используемых базисов в $\mathbf{V}_0$, $\mathbf{V}_{-1}$, $\mathbf{W}_{-1}$ дают условие на матрицы H и G:

$$H^* H + G^* G = 1$$

Обозначив вектор исходных коэффициентов через x, можно записать его разложение в сумму огрубленной версии и серии векторов деталей:

$$\begin{array}{ccccccccc} x & \xrightarrow{H} & Hx & \xrightarrow{H} & H^2 x & \xrightarrow{H} & \dots & \xrightarrow{H} & H^{N-1}x & \xrightarrow{H} & H^N x \\ \downarrow G & & \downarrow G & & \downarrow G & & & & \downarrow G & & \\ Gx & & GHx & & GH^2 x & & \dots & & GH^{N-1}x & & \end{array}$$

Эту процедуру иногда называют быстрым вейвлет-преобразованием (Fast Wavelet Transform), а иногда – алгоритмом Малла (Mallat algorithm). Число итераций N может быть произвольным. Если вектор x конечен, его надо продолжить «на бесконечность»; проще всего это сделать периодическим образом. Каждое применение операторов H и G сокращает длину вектора вдвое, поэтому общее число операций линейно по длине входа.

Численные эксперименты по распознаванию речи

Нами был разработан блок предобработки речи на основе вейвлетов.

В качестве вейвлет-базиса был использован базис Добеши-9. Поскольку этот базис является ортонормированным, то это дало возможность реализовать быстрый алгоритм вычисления вейвлет-коэффициентов на каждом частотном уровне через уже найденные коэффициенты на уровне с более высокой частотой.

Блок предобработки, который оценивает информативные параметры речевого сигнала на основе вейвлет-коэффициентов был итегрирован в систему распознавания речи ВЦ РАН [2] вместо блока вычисления мел-кепстральных параметров. Система распознавания речи основана на моделировании речевого сигнала с помощью дискретных марковских моделей аллофонов (контексто-зависимых вариантов фонем). В описанных экспериментах система распознавания включала модели для алфавита из 540 марковских моделей аллофонов, полученных с помощью построения бинарного решающего дерева. Каждая модель состояла из трех состояний. Распределение параметров речевого сигнала для каждого состояния оценивалось на материале речевого корпуса данных. Распознавание речевого сигнала выполлялось с помощью процедуры Витерби.

Эксперименты по распознаванию проводились на материале фонетической части речевого корпуса данных TeCoRus [3]. В частности использовался сигнал микрофонного качества, с частотой квантования 22050Гц. Эта частота и определила самый мелкий масштаб вейвлет-коэффициентов, соответствующий самой высокой частоте.

Применяя оконное вейвлет-преобразование, варьировалось количество частотных уровней вейвлет-коэффициентов. При заданной верхней частоте вейвлет-коэффициентов количество частотных уровней, ширина окна и количество используемых вейвлет-коэффициентов однозначно определяются нижним уровнем частот коэффициентов.

Использование слишком низкого порога нижней частоты приводит к излишнему усложнению вычислительных операций, а ограничение преобразования более высоким порогом нижних частот приводит к потерям необходимой информации для распознавания речи. С целью оптимизировать нижний уровень частот вейвлет-преобразования были проведены испытания системы для различного уровня нижних частот: 400 Гц, 800 Гц и 1600 Гц.

В данном эксперименте производились оценки точности распознавания изолированых слов: как использованных при обучении системы, так и неиспользованных на обучающей стадии. При вычислении точности распознавания учитывались произношения слов, как известными системе дикторами (чьё произношение использовалось при обучении), так и новыми для системы.

В таблице 1 приведены параметры вейвлет преобразования для используемых частот и полученная точность распознавания.

Таблица 1. Параметры вейвлет преобразования и полученная оценка ошибки распознавания.

Нижняя частота, (Гц)	Размер “окна” (точки)	Количество параметров	Ошибка распознавания (%)
400	960	25	22,6
800	480	13	9,1
1600	240	7	19,1

В следующем эксперименте использовался цифровой материал речевой части TeCoRus. Производились оценки точности для изолированного произношения и слитного произношения цифр. В таблице 2 представлены результаты тестирования системы для слитного и изолированного произношения (ошибка указана вцелом и для каждого пола отдельно).

Таблица 2. Ошибка распознавания в процентах для изолированного и слитного произношения.

Тип произношения	Мужчины	Женщины	Всего
Изолированное произношение	12	8	10
Слитное произношение	32	36	34
Всего	22	22	22

С целью повысить точность распознавания было принято решение проводить обучение системы отдельно для разных полов. В таблице 3 представлена в процентах ошибка распознавания для разных полов, на моделях, построенных для каждого пола дикторов отдельно.

Таблица 2. Ошибка распознавания в процентах для изолированного и слитного произношения. Обучение велось для каждого пола отдельно.

Тип произношения	Мужчины	Женщины
Изолированное произношение	6	6
Слитное произношение	28	36
Всего	17	21

Представленные результаты уступают результатам, которые получены на основе кепстральных параметров, однако, в данном случае эксперименты носили предварительный характер. В частности, не проводилась работа по оптимизации признаков, полученных на основе вейвлетов. Также не выполнялись некоторые, существенные для точности распознавания, процедуры предобработки (например, вычитание среднего). В настоящий момент в этих направлениях проводятся исследования. Предварительные результаты, описанные в этой статье, дают основания полагать, что после проведения оптимизации точность распознавания будет существенно улучшена, и превысит точность системы распознавания на основе кепстральных параметров.

Заключение

В работе представлены результаты работы системы распознавания речи при использовании информативных параметров, основанных на вейвлет-коэффициентах.

Благодаря получению дополнительной информации с разных масштабов времени и разных масштабов разрешения сигнала вейвлет-преобразование позволяет повысить точность и помехоустойчивость распознавания речи.

Полученные в данной работе предварительные результаты дают основания полагать, что вейвлет-анализ речевого сигнала может быть использован для построения систем распознавания изолированной и слитной речи.

Литература

1) Daubechies. Ten Lectures on Wavelets. //Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM Publications, Philadelphia, 1992.
2) Чучупал В.Я., Маковкин К.А., Чичагов А.В. К вопросу об оптимальном выборе алфавита моделей звуков русской речи для распознавания речи //Искусственный интеллект, том 4, №1, 2002, стр.575-579, Наука і освіта, Киев.
3) V.Kouznetsov, V.Chuchupal, K.Makovkin, A.Chichagov. Design and Implementation of a Russian Telephone Speech Database. //In Proc.of Int.Workshop “Speech and Computer”, Moscow, 1999, pp. 179-181.