УДК 004.934.8’1

А.Н. Ручай

Улучшение надежности текстозависимой верификации диктора на основе формантного метода с помощью нового метода сегментации речевого сигнала

Предложен новый метод сегментации речевого сигнала на вокализованные сегменты для увеличения надежности текстозависимой верификации диктора на основе формантного метода. Новый метод основан на оценке показателя сингулярности сигнала.

Ключевые слова: распознавание диктора, текстозависимая верификация диктора, форманты, ошибки первого и второго рода, сегментация речевого сигнала, сингулярность, показатель Гёльдера, непрерывное вейвлет-преобразование.

На данный момент задачу распознавания диктора нельзя считать решенной. Последние исследования в области голосовой биометрики были направлены на изучение формантного метода параметризации речевого сигнала [1–4].

В работе [4] рассматривается метод текстозависимой верификации диктора на основе анализа формантного набора, в котором речевой сигнал разбивается на сегменты. Стандартный метод покадровой обработки при этом приводит к появлению антиформант и сплошных спектров (см. [1,3]), что сказывается на точности вычисления формант и тем самым ухудшает надежность распознавания диктора.

В данной работе автором предложен новый метод для успешного разбиения речевого сигнала на непересекающиеся вокализованные сегменты. Данный метод основан на оценивании показателя сингулярности сигнала [5–7].

Сравнение метода покадровой обработки и нового предложенного метода позволяет утверждать, что новый метод сегментации позволяет уменьшить ошибки первого и второго рода в задаче текстозависимой верификации диктора.

Сегментация речевого сигнала для текстозависимой верификации диктора. В задаче текстозависимой верификации диктора на основе формантного метода формантные наборы должны вычисляться на определенных сегментах речевого сигнала.

В большинстве систем распознавания диктора используется метод покадровой обработки, в рамках которого сигнал разбивается на пересекающиеся кадры с определенной длиной и шагом смещения. Этот метод приводит к появлению провалов спектра сигнала, которые называют антиформантами, а также к сплошным спектрам [3, 8]. Вследствие данных недостатков значения формант могут быть неточными, что сказывается на надежности распознавания диктора.

Идеальным было бы вычисление формантных наборов на тех сегментах речевого сигнала, которые соответствуют фонемам, входящим в состав слова. Вследствие эффектов коартикуляции существующие методы сегментации речевого сигнала на изолированные фонетические сегменты, которые используют оценивание спектральных изменений между последовательностью речевых кадров, также могут приводить к неточным значениям формант (там же).

В статье предлагается новый метод сегментации речевого сигнала, который избавлен от появления антиформант и сплошных спектров, как в методе покадровой обработки. В новом предложенном методе сигнал сегментируется на непересекающиеся вокализованные сегменты, которые соответствуют не фонемам слова, а слогам, в основе которых лежат периодичные гласные звуки.

Предлагаемый метод заключается в оценке показателя сингулярности сигнала, в качестве которого рассматривается показатель Гёльдера. Идея использовать показатель сингулярности сигнала в качестве выделения вокализованных сегментов речевого сигнала возникла после ознакомления со статьей [5], в которой исследуется вопрос о выделении транзиентов сигнала с помощью вычисления показателя Гёльдера.

Показатель Гёльдера как оценка сингулярности сигнала. В работах [6, 8] исследуется вопрос использования оценки сингулярности сигнала с помощью показателя Гёльдера в различных практических задачах.

В силу принципа неопределенности невозможно анализировать гладкость в отдельных точках по поведению преобразования Фурье. Поэтому для оценки сингулярности сигнала используется вейвлет-преобразование.

Будем считать, что сигнал представлен функцией $f(t)$ вещественной переменной.

Согласно теореме Джаффара гладкость функции $f(t)$ в точке v характеризуется поведением непрерывного вейвлет-преобразования.

Теорема Джаффара [6]. *Пусть* ψ *– вещественный вейвлет с* n *нулевыми моментами и с быстроубывающими производными, тогда если функция* $f(t)\in \mathbf{L}^2(\mathbf{R})$ *удовлетворяет условию Гёльдера с показателем* $\alpha<n$ *в точке* v*, то существует* A *такое, что для любого* $(u,s)\in \mathbf{R}\times\mathbf{R}^+$ *будет выполнено* $|Wf(u,s)|\le As^{\alpha+1/2}\left(1+\left|(u-v)/s\right|^{\alpha}\right)$*, где* $Wf(u,s)$ *– вещественное вейвлет-преобразование функции* $f(t)$*,* s *– масштабный коэффициент.*

В частности, если $f(t)$ удовлетворяет условию Гёльдера с показателем α в точке v, и выполняется неравенство $|u-v|\le Cs$, то справедливо соотношение $|Wf(u,s)|\le A's^{\alpha+1/2}$.

Прологарифмировав обе части приведенного выше неравенства, возведенного в квадрат, получим при $u=v$ соотношение $\log\left|Wf(v,s)\right|^2\le\log A''+(2\alpha+1)\log s$, имеющее место для всех $s\in[0,+\infty)$.

Отсюда следует, что показатель α гладкости функции $f(t)$ в точке v является угловым коэффициентом опорной прямой к графику функции $\left|Wf(v,s)\right|$ в логарифмической шкале.

Сегментация речевого сигнала на основе оценки показателя сингулярности сигнала. Основная идея состоит в использовании показателя Гёльдера для разбиения речевого сигнала на вокализованные участки, которые соответствуют слогам.

В дальнейшем рассматриваем речевой сигнал как последовательность отсчетов $f(t)$, $t=1,2,..,m$, где m – число отчетов дискретизированного сигнала $f(t)$.

В качестве вейвлета выберем вещественный вейвлет Гаусса 2-го порядка.

Так как рассматривается только речевой сигнал, то в этом случае должны анализироваться частоты от 20 до 4000 Гц, причем низкие частоты 200–1500 Гц, которые соответствуют первым трем формантам, должны быть полно представлены для лучшего выделения вокализованных сегментов.

Масштабные коэффициенты в вейвлет-преобразовании выберем, пользуясь соотношением $f_S=f_W f_k/s$, где f_S – псевдочастота, s – масштабный коэффициент; f_k – частота дискретизации сигнала $f(t)$, которая равна 11025 Гц; f_W – частота центрального всплеска вейвлета, для вейвлет Гаусса 2-го порядка $f_W=0{,}3$, обоснованным в работе [9].

Тогда масштабные коэффициенты $s=1,2,...,16$ будут соответствовать псевдочастотам 200–3300 Гц и, как легко заметить, низкие частоты будут более полно представлены в псевдочастотах, которые соответствуют этим масштабным коэффициентам.

Построение опорной прямой к графику функции $\log s\to\log\left|Wf(t,s)\right|^2$ выполним следующим образом.

Для каждой точки t методом наименьших квадратов строится линейная зависимость $\hat{y}=(2\alpha(t)+1)x+\beta(t)$, минимизирующая функционал

$$Q(\alpha,\beta,X)=\sum_{s=1}^{16}((2\alpha(t)+1)x(s)+\beta(t)-y(s))^2,$$

где $(x(s),y(s))=(\log s,\log\left|Wf(t,s)\right|^2)$, $s=1,..,16$.

В качестве показателя сингулярности сигнала примем величину $\alpha(t)$.

Далее график показателя сингулярности $\alpha(t)$ сглаживается, как двумерные данные $\{(t,\alpha(t))\}_{t=1}^{m}$, с помощью метода, который был предложен Кливлендом [10]. Для полноты изложения опишем этот метод.

Введем параметр сглаживания l, $0<l<1$. Выбор параметра l обусловлен длительностью слога, вследствие экспериментов было установлено, что длительность слога в среднем равна 200 мс [3]. В качестве параметра l возьмем число, обратное количеству сегментов, которое равняется $l=f_k/5m$, где f_k – частота дискретизации сигнала $f(t)$, которая равна 11025 Гц, и m – число отчетов дискретизированного сигнала $f(t)$.

Окрестность N_t^r точки $(t,\alpha(t))$ определяется как множество индексов i, которые соответствуют ближайшим $r=\lfloor lm \rfloor$ соседним точкам $(i,\alpha(i))$ к точке $(t,\alpha(t))$ в смысле евклидового расстояния.

Для каждой точки $(t,\alpha(t))$, $t=1,..,m$, методом наименьших квадратов строится линейная зависимость $\hat{\alpha}(t)=a_t+b_t t$ по r точкам $(i,\alpha(i))$ с индексами i из окрестности N_t^r, минимизирующая следующий функционал:

$$Q(a_t,b_t,N_t^r)=\sum_{i\in N_t^r} w_i^t (a_t+b_t i-\alpha(i))^2 .$$

Здесь для каждой точки $(i,\alpha(i))$ определяется локальный вес w_i^t как

$$w_i^t=W\left(\frac{i-t}{h_t}\right),\ \text{где}\ h_t=\max_{i\in N_t^r}|i-t|$$

и, следуя [10], в качестве $W(z)$ выбирается функция

$$W(z)=\begin{cases}\left(1-|z|^3\right)^3, |z|\le 1,\\ 0, |z|>1.\end{cases}$$

Затем для каждой точки $(t,\hat{\alpha}(t))$ определим веса δ_t как $\delta_t=K\left(\hat{\varepsilon}_t/6s\right)$, где $\hat{\varepsilon}_t=|\hat{\alpha}(t)-\alpha(t)|$, s – медиана величин $\hat{\varepsilon}_1,..,\hat{\varepsilon}_m$, и

$$K(z)=\begin{cases}\left(1-|z|^2\right)^2, |z|\le 1,\\ 0, |z|>1.\end{cases}$$

Повторно для каждой точки $(t,\hat{\alpha}(t))$, $t=1,..,m$, методом наименьших квадратов строится линейная зависимость $\hat{\alpha}(t)=c_t+d_t t$ по r точкам $(i,\hat{\alpha}(i))$ с индексами i из окрестности N_t^r, минимизирующая следующий функционал:

$$Q(c_t,d_t,N_t^r)=\sum_{i\in N_t^r} w_i^t \delta_i (c_t+d_t i-\hat{\alpha}(i))^2 .$$

Полученную оценку $\hat{\hat{\alpha}}(t)$ примем за оценку показателя Гёльдера $\alpha(t)$.

На рис. 1 приведены речевой сигнал $f(t)$ и сглаженный график показателя сингулярности $\alpha(t)$ этого сигнала с параметром $l=0{,}3$. Окружностями отмечены начала вокализованных сегментов, квадратами – окончания. Было замечено, что эти выделенные точки соответствуют точкам локального минимума и максимума функции $\alpha(t)$. Интервалы монотонного возрастания показателя сингулярности $\alpha(t)$ выступают в роли вокализованных сегментов.

Для оценки качества работы предложенного метода сегментации речевого сигнала на вокализованные участки был проведен эксперимент. Для этого была собрана база голосов из 100 дикторов в возрасте от 16 до 63 лет, каждый диктор произносил 13 раз некоторое одинаковое для всех слово, содержащее 5 гласных звуков. Ко всем фразам из собранной базы был применен новый метод разбиения речевого сигнала на вокализованные сегменты с помощью оценки сингулярности сигнала. В результате были выделены вокализованные сегменты, которые полностью соответствовали 5 слогам в этом слове.

Также проводились эксперименты по сегментации с помощью предложенного метода с собранной речевой базой, состоящей из 1500 предложений длительностью 10 с дикторов в возрасте от 14 до 63 лет. Результаты экспериментов статистически подтвердили, что оценка показателя Гёльдера может быть успешно использована для сегментации речевого сигнала на вокализованные участки.

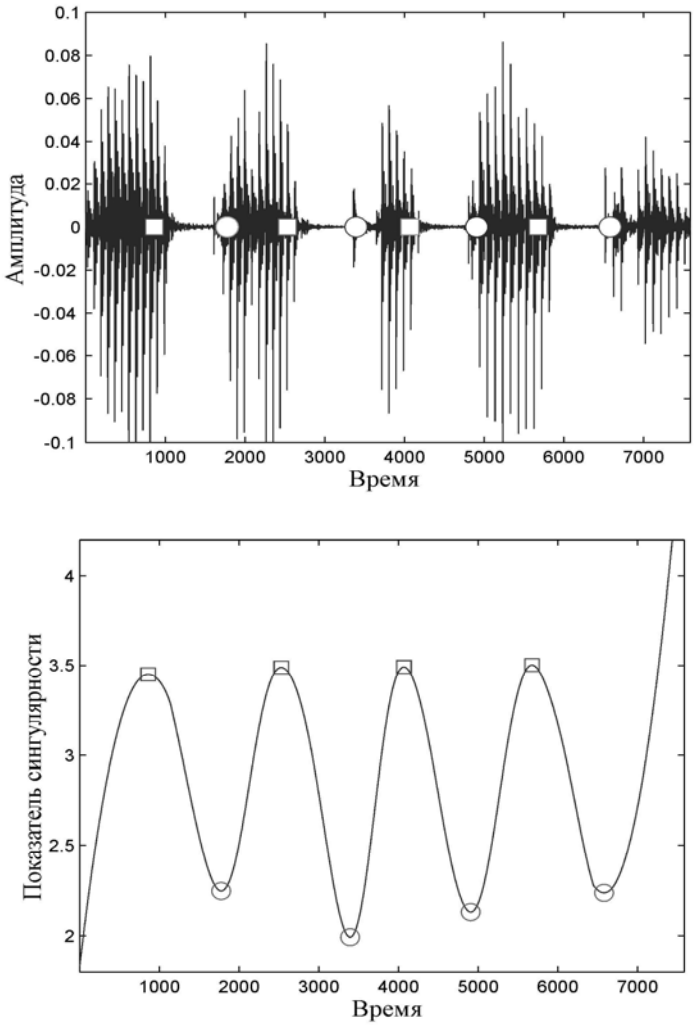

Рис. 1. Верхний рисунок – речевой сигнал, нижний – слаженный график показателя сингулярности $\alpha(t)$ для соответствующего речевого сигнала

Оценка надежности формантного метода текстозависимой верификации диктора на основе нового метода сегментации сигнала. Опишем метод получения оценки надежности текстозависимой верификации диктора на основе формантного метода с помощью покадровой обработки и нового метода сегментации сигнала, а также проведем сравнение этих оценок надежности [4].

Для этого были реализованы два метода сегментации: метод покадровой обработки и новый предложенный метод сегментации с помощью оценки сингулярности. Речевые сигналы из собранной голосовой базы разбивались на n участков. В новом предложенном методе сегментации n=5, что определяется числом слогов в слове. Для каждого сегмента вычисляем формантный набор стандартным методом [3], т.е. для каждой фразы ω находим вектор признаков $\mathbf{x}(\omega)$.

Форманту обозначим как $f=(w,a)$, где w – частота форманты, a – амплитуда форманты. Множество всех формант обозначим символом $H \subset \mathbf{R}^2$.

Под формантным набором понимаем набор формант $F=\{f_i\}_{i=1}^{u}=\{(w_i,a_i)\}_{i=1}^{u}$, где $u \in \mathbf{N}$ и $w_i < w_j$, если $i<j$. Множество всевозможных формантных наборов обозначим как V.

Сравнение фраз дикторов ω_i и ω_j осуществляется при помощи решающего правила при заданном пороговом значении λ:

$$\hat{g}(\mathbf{x}(\omega_i),\mathbf{x}(\omega_j))=\begin{cases}1, & \text{если } S(\omega_i,\omega_j)<\lambda;\\ 0, & \text{иначе,}\end{cases} \quad (1)$$

где ω_i и ω_j – объекты распознавания, соответствующие i-й и j-й фразе дикторов. Меру близости $S(\omega_i,\omega_j)$ определим как $S(\omega_i,\omega_j)=\frac{1}{n}\sum_{t=1}^{n}h(x_i^t,x_j^t)$, где $\mathbf{x}_i$ и $\mathbf{x}_j$ – векторы признаков объектов ω_i и ω_j (по числу сегментов n). Метрику в пространстве формантных наборов V введем соотношением $h(x_i^t,x_j^t)=\frac{1}{u}\sum_{l=1}^{u}r(f_{il}^t,f_{jl}^t)$, где $x_i^t=\{f_{il}^t\}_{l=1}^u$ и $x_j^t=\{f_{jl}^t\}_{l=1}^u$ – формантные наборы для t-й координаты векторов признаков $\mathbf{x}_i$ и $\mathbf{x}_j$, $u=8$ как самое оптимальное значение [4]. Здесь $r(f_{il}^t,f_{jl}^t)=c_w|w_{il}^t-w_{il}^t|+c_a|a_{il}^t-a_{il}^t|$ – метрика в пространстве формант H с весовыми коэффициентами c_w и c_a, которые определяют допустимый предел порогового значения λ, с формантами $f_{il}^t=(w_{il}^t,a_{il}^t)$ и $f_{il}^t=(w_{il}^t,a_{il}^t)$ в формантных наборах x_i^t и x_j^t. Следуя рекомендации в [4], для нормировки выберем $c_w=1$ и $c_a=1000$.

Чтобы получить количественную оценку надежности текстозависимой верификации диктора на основе формантного метода, необходимо найти ошибки первого и второго рода.

С этой целью по всевозможным фразам дикторов из собранной голосовой базы составим матрицу $M_{1300\times1300}$, элементами которой являются 1 или 0, соответствующие результатам решающего правила (1).

При успешном распознавании дикторов в идеальном случае матрица M должна содержать единицу только в тех местах, где фразы соответствуют одному и тому же диктору. Значит, количество единиц для такой матрицы должно быть равно $c_1=1300\cdot13$, а нулей должно быть $c_0=1300\cdot(1300-13)$. Стоит отметить, что матрица M является симметричной.

В построенной матрице M, в тех местах, где фразы соответствуют одному и тому же диктору, подсчитываем количество нулей d_0. А в тех местах, где фразы соответствуют разным дикторам, подсчитываем количество единиц d_1. Ошибки первого p_1 и второго p_2 рода введем следующим образом: $p_1=d_0/c_1$ и $p_2=d_1/c_0$.

Перебирая различные пороговые значения λ, вычисляем ошибки первого рода $p_1(\lambda)$ и второго рода $p_2(\lambda)$ для этих пороговых значений. Для того чтобы сравнить надежность систем распознавания диктора на основе двух методов сегментации сигнала, фиксируем ошибку второго рода $p_2(\lambda')\approx0{,}01$ и сравниваем ошибки первого рода при полученном пороге λ'. В таблице приведены ошибки первого рода $p_1(\lambda')$ и второго рода $p_2(\lambda')\approx0{,}01$.

Если сравнивать оценки надежности распознавания диктора на основе метода покадровой обработки и нового предложенного метода сегментации сигнала, то можно сделать вывод, что ошибка первого рода уменьшилась на 20% при фиксированной ошибке второго рода.

Результаты сравнения двух методов сегментации речевого сигнала

Метод сегментации	Ошибка 1 рода	Ошибка 2 рода
Покадровая обработка	0,377	0,01
Оценка сингулярности сигнала	0,301	0,01

Выводы. На основании результатов экспериментов можно утверждать, что предложенный новый метод успешно разбивает речевой сигнал на непересекающиеся вокализованные сегменты и может быть применен к различным задачам.

Одной из таких задач является текстозависимая верификация диктора на основе формантного метода. Из экспериментов было установлено, что с помощью предложенного метода сегментации сигнала ошибка первого рода уменьшилась на 20% при фиксированной ошибке второго рода для распознавания диктора по сравнению со стандартным методом покадровой обработки.

Также стоит отметить, что предложенный метод сегментации может быть использован для выделения участков, содержащих отдельные фонемы, что требует дальнейших исследований.

Литература

1. Рамишвили Г.С. Автоматическое опознавание говорящего по голосу. – М.: Радио и связь, 1981. – 224 с.
2. Репалов С.А. Разработка математических моделей и робастных алгоритмов идентификации дикторов по их речи: дис. … канд. физ.-мат. наук: 05.13.18 / С.А. Репалов. – Ростов-на-Дону, 2003. – 140 с.
3. Аграновский А.В. Теоретические аспекты алгоритмов обработки и классификации сигналов / А.В. Аграновский, Д.А. Леднов. – М.: Радио и связь, 2004. – 164 с.
4. Ручай А.Н. Формантный метод текстозависимой верификации диктора // Вестник Челяб. гос. университет. Математика. Механика. Информатика. – 2010. – №23(204), вып. 12. – С. 121–131.
5. Хабибуллин Р.Ф. Локализация транзиентов в звуковых сигналах с помощью оценки локального показателя Гёльдера / Р.Ф. Хабибуллин, Л.И. Левкович-Маслюк / Препринт Института прикладной математики им. М.В. Келдыша РАН. – М., 2006.
6. Малла С. Вейвлеты в обработке сигнала. – М.: Мир, 2005. – 671 с.
7. Ручай А.Н. Текстозависимая верификация диктора на основе формантного метода с использованием нового метода сегментации речевого сигнала // Современные проблемы математики: тезисы 42-й Всерос. молод. конф. – Екатеринбург: УрО РАН, 2011. – С. 164–166.
8. Айфичер Э. Цифровая обработка сигналов: практический подход / Э. Айфичер, Б. Джервис. – М.: Вильямс, 2004. – 992 с.
9. Abry P. Ondelettes et turbulence. Multirésolutions, algorithmes de décomposition, invariance d'échelles. – Paris: Diderot Editeur, 1997.
10. Cleveland W.S. Smoothing by local regression: principles and methods / W.S. Cleveland, C.L. Loader // Statistical Theory and Computational Aspects of Smoothing. – New York: Springer, 1996. – P. 10–49.

Ручай Алексей Николаевич
Аспирант каф. компьютерной безопасности и прикладной алгебры ЧелГУ, г. Челябинск
Тел.: 8 (351) 977-92-92
Эл. почта: ruchai@pochta.ru

Ruchay A.N.

Improvement in safety of formant method of text dependent verification of a speaker by a new method of segmentation of speech signal

The new method of segmentation of speech signal is introduced in this article, it improves the safety of the formant method of text dependent verification of a speaker. The new method is based on an exponent of signal singularity.

Keywords: speaker recognition, text dependent speaker verification, formant, false acceptance and rejection rate, segmentation of speech signal, singularity, Holder exponent, continuous wavelet transform.