УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
Том 155, кн. 2 Физико-математические науки *2013*

УДК 534.4+004.93

ПАРАМЕТРЫ, ХАРАКТЕРИЗУЮЩИЕ ЛОКАЛЬНЫЕ ФРАГМЕНТЫ РЕЧЕВЫХ ФАЙЛОВ

Р.Р. Нигматуллин, Е.Л. Столов

Аннотация

Предложены параметры и способы их оценки для описания участков звукового файла, соответствующих отдельным слогам. Для слогов, начинающихся со взрывной согласной, указан метод однозначной локализации «вспышки». Для описания участка огласовки предложен способ оценки мгновенной частоты, не требующий применения преобразования Гильберта. Показано, что распределение мгновенных частот в речевом файле может служить одним из параметров для идентификации диктора.

Ключевые слова: взрывные согласные, локализация вспышки, аппроксимация мгновенной частоты, распределение мгновенных частот, идентификация диктора.

Введение

В настоящее время в связи с появлением в свободном доступе большого количества речевых файлов возникла задача автоматического определения языка, использованного при создании конкретного файла. Предполагается, что файл задан в одном из стандартных цифровых форматов. Требуется определить язык из заданного списка, использованный для записи речи. Решение достигается путем выбора параметров, измеряемых по данному файлу, и способа их оценки. В последнее время был предложен ряд подходов для решения указанной задачи. Самый простой метод предполагает подсчет отдельных статистик по отрезкам речевого файла фиксированной длины. Стандартная процедура на основе метода главных компонентов применяется для описания указанных отрезков [1]. В [2] для идентификации языка использован тот же подход, но на последнем шаге классификация осуществлялась с помощью метода опорных векторов. Указанный подход позволяет провести классификацию заданного числа звуковых файлов, но не объясняет природу особенностей того или иного языка. Более перспективными представляются подходы, когда анализируются особенности фонем, присутствующих в звуковом файле [3], используются спектральные свойств сигнала и проводится дальнейшая классификация с помощью GMM (Gaussian Mixture Model) [4]. Отдельно стоит метод идентификации языка на основе визуальной информации – анализа движения губ говорящего [5]. В настоящей работе принят подход, согласно которому исследуются фонемы специального вида – слоги, начинающиеся с глухих взрывных согласных (п, к, т), за которыми следует гласная. Рассмотрим типичную оцифровку фрагмента звукового файла, относящегося к фонеме ‘ka’. Фрагмент представлен на рис. 1. Он состоит из следующих трех частей.

1. Участок, возникший в результате быстрого раскрытия связок, который при прослушивании воспринимается как резкое изменение интенсивности сигнала. Этот участок принято называть вспышкой. На рис. 1 он выделен стрелками.
2. Шум, который расположен между вспышкой и началом огласовки.
3. Огласовка последующей гласной.

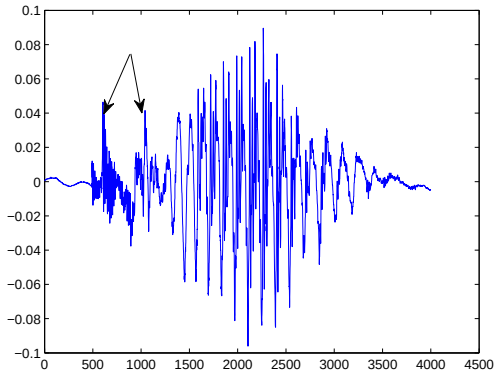

Рис. 1. Положение «вспышек» в фонеме ‘ka’

В [6] показано, как можно описать форму сигнала, отвечающего вспышке. Там же установлено, что форма вспышки может служить признаком, с помощью которого осуществляется идентификация диктора из небольшого списка, заданного заранее. Известно [7], что для тюркских языков характерно укороченное расстояние между вспышкой, отвечающей согласной, и началом последующей огласовки, поэтому измерение данного параметра является важным шагом на пути определения языка.

Определение начала огласовки оказалось непростой задачей. Обычно это делается оператором путем прослушивания соответствующего фрагмента. В [8] предложен алгоритм для определения начала огласовки, основанный на кластеризации коэффициентов Фурье.

1. Локализация вспышки

При практических измерениях нужно иметь объективный способ локализации вспышки, не зависящий от оператора. Задача усложняется тем, что обычно сигнал имеет сложную форму. Часто для этой цели предлагается использовать положение максимума. Как следует из рис. 1, этих положений может быть несколько. В статье предлагается метод для однозначного определения позиций вспышки, не зависящий от формы сигнала. В основе предлагаемого подхода лежит следующая простая

Теорема 1. *Пусть даны две неотрицательные четные функции* $f_1(x)$, $x \in [-b, b]$ *и* $f_2(x)$, $x \in [-a, a]$, $a \le b$, f_2 *ограничена,* f_1 *дважды непрерывно дифференцируема и выпукла, то есть* $f_1''(x) < 0$. *Тогда функция*

$$F(p) = \int_{-a}^{a} f_1(x+p) f_2(x)\, dx \tag{1}$$

имеет локальный максимум при $p = 0$.

Доказательство. В силу четности функции f_1 имеет место равенство $f_1'(-x) = -f_1'(x)$. В окрестности $p = 0$ получаем

$$f_1(x+p) = f_1(x) + p f_1'(x) + p^2 f_1''(x)/2 + p^3 O(1),$$

поэтому

$$F(p) = \int_{-a}^{a} (f_1(x)f_2(x) + pf_1'(x)f_2(x) + p^2\frac{f_1''(x)}{2}f_2(x))\,dx + p^3O(1). \qquad (2)$$

В силу сделанных предположений функция $f_1'f_2$ является нечетной, поэтому второе слагаемое в (2) под знаком интеграла обращается в 0, третье слагаемое при $p \neq 0$ будет отрицательным. □

Теорема 1 применяется для локализации вспышки в звуковом файле следующим образом.

1. Выделяется интервал $[-b, b]$, внутри которого находится вспышка, и сигнал внутри интервала считается заданным с помощью функции f_2.

2. Все отрицательные значения функции заменяются нулями.

3. Выбирается функция f_1 и находятся положения локальных максимумов функции (1).

Найденные положения максимумов будут одними и теми же для любой функции f_1, удовлетворяющей условиям теоремы 1. Простейшей из них является функция, определенная равенством $f_1(x) = -x^2 + b^2$ на интервале $[-a, a]$ и равная нулю вне этого интервала. Именно с этой функцией производятся дальнейшие вычисления. С другой стороны, функции f_2 на практике лишь приближенно могут считаться четными. В этой связи возникает вопрос, каким образом отсутствие симметрии влияет на положение максимума. Представим функцию f_2 в виде суммы симметричной и кососимметричной функций

$$f_2(x) = (f_2(x) + f_2(-x))/2 + (f_2(x) - f_2(-x))/2 = s(x) + c(x).$$

Равенство (2) с учетом симметричности интервала интегрирования принимает форму

$$F(p) = \int_{-a}^{a} (-(x+p)^2 + b^2)f_2(x)\,dx = -\int_{-a}^{a} ((x^2 + p^2 + b^2)s(x) + 2pxc(x))\,dx.$$

Условие $F'(p) = 0$ приводит к соотношению

$$p = -\int_{-a}^{a} xc(x)\,dx \Big/ \int_{-a}^{a} s(x)\,dx. \qquad (3)$$

Равенство (3) указывает, на сколько будет смещено относительно нуля положение максимума при наличии кососимметрической составляющей в функции f_2.

Доказанная теорема 1 указывает, каким образом можно локализовать положение вспышки в звуковом файле, когда имеют дело с функцией, заданной для дискретных значений аргумента. Сначала заменим все значения в файле, не превышающие некоторый порог (уровень шума), нулями и обозначим получившуюся функцию через f_2. Заменим интеграл в (1) суммой, в результате вычисление сводится к применению FIR-фильтра с коэффициентами, определенными значениями $f_1(x) = -x^2 + b^2$ внутри интервала $[-a, a]$, к значениям функции f_2. Параметр a подбирается таким образом, чтобы предполагаемая длина интервала вспышки была меньше $2a$. Результат применения фильтра к исследуемому участку представлен на рис. 2. Здесь четко выделены положения локальных максимумов, которые и объявляются позициями вспышек.

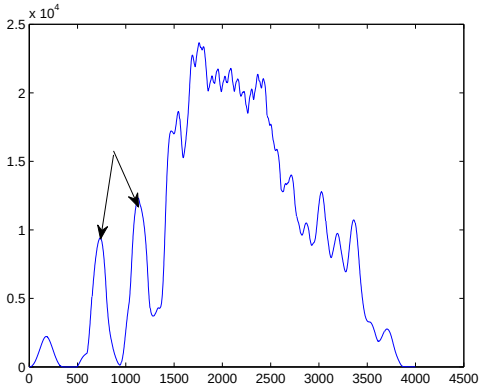

Рис. 2. Результат фильтрации участка файла с фонемой 'ka'

2. Аппроксимация мгновенной частоты сигнала

Еще более трудной задачей является описание участка файла, относящегося к огласовке. Как правило, для этой цели применяют кратковременное преобразование Фурье [4, 8]. Однако при описании переходного периода с помощью этого преобразования возникают очевидная трудность – сигнал не будет стационарным, результат зависит от длины окна, поэтому найденные частоты и их магнитуды носят усредненный характер. В этой связи представляет интерес подсчет мгновенных частот в некоторых точках сигнала, отвечающих огласовке. Напомним основные определения, связанные с понятием мгновенной частоты [9]. Пусть имеется сигнал $u(t)$. Строится комплексный сигнал $z(t)$ такой, что его спектр содержит только положительные значения, и $u(t) = \operatorname{Re} z(t)$. Сигнал $z(t)$ называется аналитическим. Для него существует представление $z(t) = r(t)\exp(i\theta(t))$, где $r(t)$ – модуль $z(t)$, а производная $\theta'(t)$ называется мгновенной частотой сигнала в момент времени t. Мгновенная частота является важной характеристикой сигнала, однако вычисление $z(t)$ предполагает применение преобразования Гильберта к исходному сигналу $u(t)$. Это преобразование определяется формулой

$$w(t) = \frac{1}{\pi}\int\limits_{-\infty}^{\infty} \frac{u(x)}{t-x}\,dx, \tag{4}$$

а $z(t) = u(t) + iw(t)$. В случае функции, заданной для дискретных значений аргумента, это преобразование реализовано в виде фильтра специального вида. Фильтр не является физически реализуемым, поскольку определение значения сигнала в момент времени t, как и в случае (4), предполагает знание исходного сигнала в моменты времени, как предшествующие моменту t, так и после этого момента. На практике при вычислениях используют лишь значения сигнала на некотором интервале, но при этом предполагается стационарность сигнала на этом интервале. К сожалению, время огласовки не удовлетворяет этому ограничению, поэтому полученные стандартным образом значения трудно интерпретировать. В настоящей работе отмеченная проблема преодолевается путем аппроксимации исходного сигнала другим, для которого известен явный вид преобразования Гильберта.

Преобразование Гильберта от сигнала $u(t) = \cos(wt)$ есть $\sin(wt)$, а преобразование от $u(t) = \sin(wt)$ равно $-\cos(wt)$ [9]. Мгновенная частота сигнала

$$x(t) = A\cos(wt) + B\sin(wt) \tag{5}$$

в любой точке равняется w и не зависит от коэффициентов A, B. Пусть дана дискретная версия $u[n]$ непрерывного исследуемого сигнала $u(t)$, полученная согласно формуле $u[n] = u(nT)$, где T – некоторый интервал времени, $f = 1/T$. Предположим, что сигнал $u[n]$ имеет положительный локальный максимум при $n = n_0$. Без ограничения общности можем предполагать, что $n_0 = 0$. Приблизим в окрестности этой точки сигнал $u(t)$ сигналом вида (5). Имеем

$$A = u[0], \quad A\cos(w/f) + B\sin(w/f) = u[1], \quad A\cos(w/f) - B\sin(w/f) = u[-1].$$

Отсюда следует, что

$$\cos(w/f) = \frac{u[1] + u[-1]}{2u[0]}. \tag{6}$$

Из сделанных предположений о существовании локального максимума вытекает, что правая часть в (6) меньше 1, и определение частоты становится корректным.

Согласно (6) оценка мгновенной частоты в точке локального максимума $u[0]$ зависит от трех последовательных отсчетов функции, но сами эти значения зависят от начала отсчета. Покажем, что в случае, когда функция $u(t)$ имеет вид (5), результат, подсчитанный по формуле (6), от начала отсчета не зависит. Действительно, пусть

$$\begin{aligned}
u[0] &= A_1\cos(wn/f) + B_1\sin(wn/f),\\
u[1] &= A_1\cos(w(n+1)/f) + B_1\sin(w(n+1)/f),\\
u[-1] &= A_1\cos(w(n-1)/f) + B_1\sin(w(n-1)/f).
\end{aligned}$$

Непосредственными вычислениями легко убедиться, что в этом случае формула (6) превращается в тождество.

В этой связи интересно посмотреть, какие значения получаются в результате предложенной вычислительной процедуры для сигнала с известной мгновенной частотой. Предположим, что исходный сигнал имеет вид

$$u(t) = \sum_k (A_k\cos(w_k t) + B_k\sin(w_k t)). \tag{7}$$

Преобразование Гильберта от этой функции равно

$$v(t) = \sum_k (A_k\sin(w_k t) - B_k\cos(w_k t)).$$

Из определения аналитического сигнала следует, что

$$\operatorname{tg}(\theta(t)) = \frac{v(t)}{u(t)}.$$

Найдем мгновенную частоту сигнала

$$(1 + \operatorname{tg}^2(\theta(t))\theta'(t) = \frac{v'(t)u(t) - u'(t)v(t)}{u^2(t)}.$$

При $t = 0$ мгновенная частота есть

$$\theta'(0) = \frac{\Big(\sum\limits_k w_k A_k\Big)\Big(\sum\limits_k A_k\Big) + \Big(\sum\limits_k w_k B_k\Big)\Big(\sum\limits_k B_k\Big)}{\Big(\sum\limits_k A_k\Big)^2 + \Big(\sum\limits_k B_k\Big)^2}. \tag{8}$$

Табл. 1

Сравнение точных значений мгновенной частоты с оценкой

Формула (8)	3729	3517	3788	36261	3418	3510	3152	3784	3879	3546
Формула (9)	3835	3572	3800	3676	3438	3552	3016	4047	3454	3680

С другой стороны, применив к сигналу $u(t)$ формулу (6), получим

$$\cos(w/f) = \frac{\sum\limits_k A_k \cos(w_k/f)}{\sum\limits_k A_k}, \tag{9}$$

то есть при положительных коэффициентах A_k значение $\cos(w/f)$ есть выпуклая комбинация косинусов исходного сигнала. Для сравнения значений мгновенных частот, полученных согласно (8) и вычисленных по формуле (9), случайным образом восемь раз порождались коэффициенты A_k, B_k, а частоты w_k также выбирались случайным образом, но только один раз. Случайный выбор коэффициентов A_k, B_k равносилен оценке мгновенной частоты в разных точках. Ниже приведены результаты экспериментов, когда число слагаемых в (7) равнялось 5, а в (9) положено $f = 16000$.

Как следует из табл. 1, обе строки содержат числа одного порядка.

3. Идентификация диктора с помощью распределения мгновенных частот

В настоящее время считается решенной задача идентификации диктора по речевому сигналу, если продолжительность этого сигнала превышает минуту. В основе процедуры лежит достаточно сложная техника, предложенная в [10]. С тех пор появилось большое количество работ, в которых алгоритм модифицировался, но все эти работы использовали GMM в качестве основы. Утверждается, что с помощью этого метода диктор идентифицируется правильно в 95% случаев, однако сама процедура идентификации требует значительных ресурсов. В настоящей работе не решалась задача идентификации диктора, но было обнаружено, что распределение мгновенных частоты, полученных с помощью описанной выше процедуры, позволяет идентифицировать диктора достаточно точно. Оценкам достоверности метода будет посвящена следующая работа, а пока ограничимся лишь одним примером.

Процедура сравнения двух речевых файлов проводилась следующим образом.

1. Определялся уровень шума, и все сигналы ниже выбранного уровня полагались равным нулю. При вычислениях порог полагался равным 0.3 от стандартного отклонения.

2. В преобразованном файле выделялись точки локального максимума и в каждой из этих точек вычислялась мгновенная частота согласно (6).

3. Строилась гистограмма распределения частот, порождающая выборочную функцию распределения.

4. Расстояние между двумя выборочными распределениями d_1, d_2 определялось согласно формуле $\mathrm{dist} = \max(|\mathrm{d1} - \mathrm{d2}|)$ (распределение Колмогорова–Смирнова).

На рис. 3 представлен пример типичной гистограммы. Табл. 2, 3, содержащие результаты измерения расстояний между речевыми файлами, сформированы следующим образом. Взяты два речевых файла объемом по 10^7 отсчетов каждый, принадлежащих двум дикторам. Файлы записаны с частотой 44100 Гц в студии. Файл разбивался на 5 частей, и каждая часть обрабатывалась независимо. В табл. 2 приведены расстояния между частями, принадлежащими одному диктору. Через

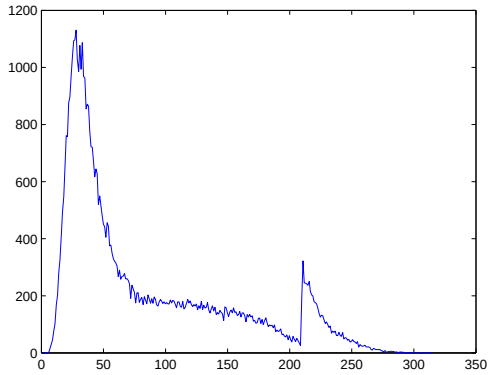

Рис. 3. Пример гистограммы распределения мгновенный частот

Табл. 2

Сравнение расстояний между частями речевого файла, принадлежащих одному диктору

$S(x,y)=$

0	0.0315	0.0242	0.0206	0.0203
0.0224	0	0.0234	0.0175	0.0173
0.0807	0.0788	0	0.0130	0.0119
0.0187	0.0177	0.0646	0	0.0069
0.0248	0.0244	0.0811	0.0205	0

Табл. 3

Сравнение расстояний между частями речевого файла, принадлежащих двум разным дикторам

$T(x,y)=$

0.1065	0.1217	0.1146	0.1073	0.1128
0.1186	0.1319	0.1266	0.1177	0.1238
0.0943	0.0652	0.0844	0.0755	0.0817
0.1061	0.1195	0.1142	0.1054	0.1113
0.1060	0.1214	0.1151	0.1066	0.1121

$S(x,y)$, $x < y$ обозначается расстояние между частями с номерами x и y для первого диктора, а $S(x,y)$, $x > y$ аналогичные значения для второго диктора.

В табл. 3 через $T(x,y)$ обозначено расстояние между двумя частями речевых файлов, принадлежащих разным дикторам. Анализ приведенных в табл. 2, 3 данных показывает, что расстояния между частями, принадлежащими одному диктору существенно меньше расстояний между частями речевых файлов, принадлежащим разным дикторам. Данное обстоятельство позволяет надеяться, что предлагаемая процедура может быть использована для идентификации диктора, поскольку объем требуемых вычислений оказывается минимальным.

Работа выполнена при финансовой поддержке РФФИ (проект № 12-01-97002-р-поволжье-а).

Summary

R.R. Nigmatullin, E.L. Stolov. Parameters Describing Local Properties of Speech Records.

We suggest some new parameters for the description of short fragments of speech signals as well as methods for their estimation. We develop a technique for the exact localization

of "explosion" in syllables beginning with stop consonants. A method for the evaluation of instantaneous frequency in the sound part of the syllable is presented. The method does not suppose the implementation of the Hilbert transform. It is shown that the distribution of instantaneous frequencies in a speech signal can be used for speaker identification.

Keywords: stop consonants, explosion localization, approximation of instantaneous frequency, distribution of instantaneous frequencies, speaker identification.

Литература

1. *Li H., Ma B., Lee C.-H.* A vector space modeling approach to spoken language identification // IEEE Audio, Speech, Language Process. – 2007. – V. 15, No 1. – P. 271–284.
2. *Campbell W.M., Campbell J.P., Reynolds D.A., Singer E., Torres-Carrasquillo P.A.* Support vector machines for speaker and language recognition // Comput. Speech Lang. – 2006. –V. 20, No 2–3. – P. 210–229.
3. *Siniscalchi S.M., Reed J., Svendsen T., Lee C.-H.* Universal attribute characterization of spoken languages for automatic spoken language recognition // Comput. Speech Lang. – 2013. – V. 27, No 1. – P. 209–227.
4. *Koolagudi S.G., Rastogi D., Rao K.S.* Spoken language identification using spectral features // Commun. Comput. Inform. Sci. – 2012. – V. 306. – P. 496–497.
5. *Newman J.L., Cox S.J.* Language identification using visual features // IEEE Audio, Speech, Language Process. – 2012. – V. 20, No 7. – P. 1936–1947.
6. *Нигматуллин Р.Р., Столов Е.Л.* Различение двух дикторов по коротким фразам неортогональным вейвлет преобразованием // Исслед. по прикл. матем. и информатике. – Казань: Изд-во Казан. ун-та, 2011. – Вып. 27. – С. 153–160.
7. *Diehl R.L., Lotto A.J., Holt L.L.* Speech perception // Annu. Rev. Psychol. – 2004. – V. 55. – P. 149–179.
8. *Нигматуллин Р.Р., Столов Е.Л.* Определение времени установления вокализации в слогах, начинающихся с глухой согласной // Вестн. КГТУ им. А.Н. Туполева. – 2011. – № 1. – С. 159–163.
9. *Лайонс Р.* Цифровая обработка сигналов. – М.: Бином, 2006. – 652 с.
10. *Reynolds D.A., Rose R.C.* Robust text independent speaker identification using Gaussian mixture speaker models // IEEE Speech Audio Process. – 1995. – V. 3, No 1. – P. 72–83.

Поступила в редакцию
21.12.12

Нигматуллин Руслан Рафикович – аспирант кафедры системного анализа и информационных технологий, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *nigmatullin.ruslan@gmail.com*

Столов Евгений Львович – доктор технических наук, профессор кафедры системного анализа и информационных технологий, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *ystolov@kpfu.ru*