Математические структуры и моделирование
2011, вып. 24, с. 12–18

УДК 004.934

ПРИМЕНЕНИЕ ПРЕОБРАЗОВАНИЯ ГИЛЬБЕРТА-ХУАНГА К ЗАДАЧЕ СЕГМЕНТАЦИИ РЕЧИ

О.А. Вишнякова, Д.Н. Лавров

В данной статье представлен новый подход к решению задачи сегментации речи, основанный на анализе нестационарных и нелинейных сигналов. Границы сегментов устанавливаются на участках быстрого изменения модовых функций. Приведено сравнение с результатами ручной разметки.

Введение

В системах обработки речи (распознавание речи, компрессия речи, идентификация диктора по голосу и т. д.) первоочередной задачей, требующей решения, является задача сегментации. Под сегментацией понимается операция разбиения речи на лингвистические элементы [1].

Эти основные лингвистические элементы называют фонемами, а их часто разнообразные, различимые варианты — аллофонами. Фонемы можно рассматривать как некоторый код, однозначно связанный с артикуляторными движениями данного языка. Аллофоны же данной фонемы представляют собой как бы акустическую степень свободы в реализации кодового символа. Свобода в реализации зависит не только от самой фонемы, но и от её положения во фразе [6].

В русском языке насчитывают 41 фонему, в том числе 7 гласных (а, о, у, э, и, й, ы) и 34 согласных. Согласные звуки подразделяются на звонкие согласные (з, зь, ж, м, мь, н, нь, л, ль, р, рь, в, вь), глухие согласные (ф, фь, с, сь, ш, х, хь, ц, чь) и взрывные согласные (б, бь, п, пь, д, дь, т, ть, г, гь, к, кь). Длительность звуков речи изменяется в пределах от 25 до 250 мс [1].

Традиционными подходами к задаче сегментации являются методы, основанные на Фурье, вейвлет-анализе, а также на методе линейного предсказания. Однако они не учитывают природу человеческой речи, а именно её нестационарность и нелинейность. В статье предложен новый подход к задаче сегментации, общий алгоритм сегментации и первые результаты работы алгоритма.

Омский государственный университет им. Ф.М. Достоевского
E-mail: olga@infotekorg.ru, lavrov@omsu.ru

Таблица 1. Соотношения дисперсии звуков и D_λ

Гласные звуки	Звонкие согласные	Глухие согласные	Взрывные согласные
$D_{\text{гл.}} = 2D_\lambda$	$D_{\text{з.с.}} = D_\lambda$	$D_{\text{г.с.}} = 0{,}01D_\lambda$	$D_{\text{в.с.}} < D_\lambda$

1. Нестационарность речевого сигнала

Рассмотрим речь как непрерывный случайный процесс $\lambda(t)$, образованный следующими друг за другом звуками. Каждый звук, в свою очередь, представляет реализацию случайного процесса с различными вероятностными характеристиками [3]. Оценим статистические характеристики речевого сигнала.

Гласные звуки образуются посредством возбуждения голосового тракта колебаниями голосовых связок. При нормальной артикуляции речевой тракт сохраняет относительно стабильную конфигурацию на большей части протяжённости звука. Гласные звуки произносятся с открытым ртом и обладают большой дисперсией $D_{\text{гл}}$.

Согласные звуки, как правило, характеризуются более узкой артикуляционной щелью, чем гласные звуки. Звонкие согласные, как и гласные, образуются при колебании голосовых связок, но т.к. произносятся с полуоткрытым ртом, то обладают меньшей дисперсией $D_{\text{з.с.}}$, чем гласные. При произнесении глухих согласных голосовые связки не участвуют, звуки формируются шумом выдыхаемого воздуха и обладают небольшой дисперсией $D_{\text{г.с.}}$.

Взрывной звук формируется речевым аппаратом человека в виде паузы длительностью порядка 100 мс с последующим быстрым, характерным для данного взрывного звука нарастанием последующего гласного звука. Так как большую часть взрывного согласного звука составляет пауза, то его дисперсия $D_{\text{в.с.}}$ весьма мала.

Обозначим через D_λ среднее значение дисперсии речевого сообщения $\lambda(t)$. В таблице 1 представлены соотношения дисперсии звуков и D_λ.

При статистическом описании речевого процесса полагают, что ввиду чередования звуков речь представляет собой нестационарный случайный процесс с изменяющейся дисперсией. Плотность вероятности отдельных звуков близка к нормальной, и поэтому речевое сообщение $\lambda(t)$ в фиксированный момент времени определяется нормальной нестационарной плотностью вероятности

$$f(\lambda, t) = \frac{e^{\frac{\lambda^2}{2D_\lambda(t)}}}{\sqrt{2\pi D_\lambda(t)}}.$$

Усреднённая по времени нормированная спектральная плотность $S_\lambda(f)$ речевого процесса, полученная экспериментальным путём, приведена на рисунке 1, частота по оси абсцисс отложена в логарифмическом масштабе.

Максимум $S_\lambda(f)$ соответствует частоте, близкой к $f = 500$ Гц. Ширина спектра по уровню $0{,}5$ составляет около 300 Гц. Положение максимума в реальных условиях может лежать в пределах от 400 до 600 Гц в зависимости от дисперсии речевого процесса $D_\lambda(t)$. С увеличением частоты спектральная

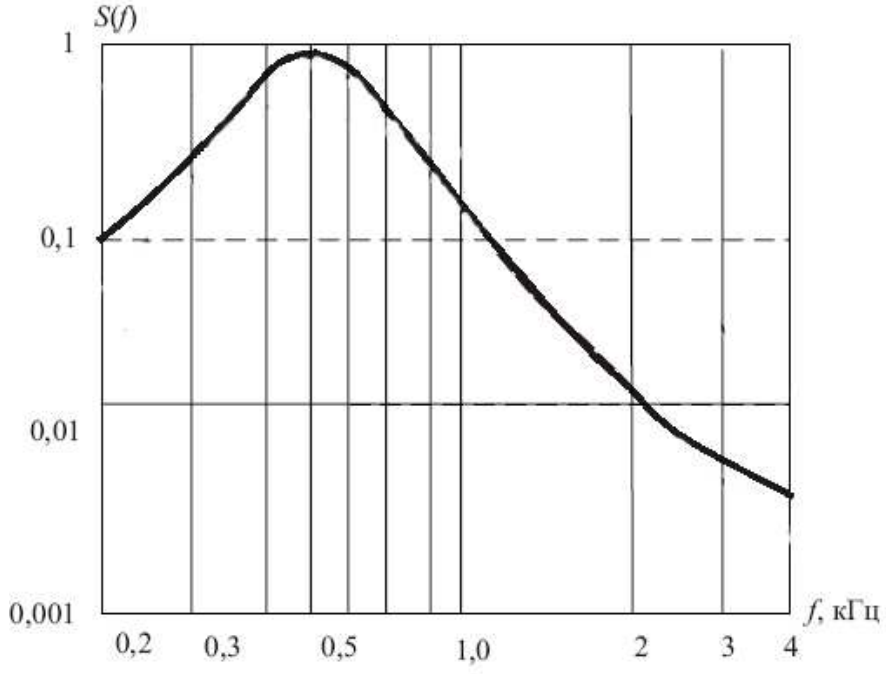

Рис. 1. Средняя нормированная спектральная плотность

плотность уменьшается. Однако, несмотря на малую интенсивность высокочастотных составляющих $S_\lambda(f)$, они оказывают значительное влияние на разборчивость речи [4].

2. Нелинейность речевого сигнала

С физической и математической точек зрения в традиционном линейном подходе к моделированию речевого сигнала реальные нелинейные процессы речеобразования аппроксимируются с использованием стандартного предположения о линейности акустических характеристик. Акустическая система при этом приближённо может быть описана одномерным волновым уравнением, например, уравнением Вебстера, в котором предполагается синфазное расположение фронтов волны по площади поперечного сечения [6]:

$$\frac{1}{A(x)}\frac{\partial}{\partial x}\left[A(x)\frac{\partial p}{\partial x}\right] = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2},$$

где $A(x)$ — площадь поперечного сечения речевого тракта, p — звуковое давление, c — скорость распространения звука. Несмотря на успешное применение линейной модели в некоторых прикладных задачах, нельзя не учитывать теоретические и экспериментальные доказательства существования явлений нелинейной трехмерной нестабильной динамики в процессе речеобразования, которые не могут быть описаны линейной моделью. Примерами таких явлений могут служить модуляция речевого воздушного потока и турбулентность.

В работе [7] представлены некоторые физические измерения, показывающие турбулентность в воздушном потоке.

Существует несколько аргументов, подтверждающих явления нестационарности в речевом сигнале [8].

3. Метод анализа нестационарных и нелинейных сигналов

Под преобразованием Гильберта-Хуанга (HHT) понимается метод эмпирической модовой декомпозиции (EMD) нелинейных и нестационарных процессов и Гильбертов спектральный анализ (HSA). HHT представляет собой частотно-временной анализ данных и не требует априорного функционального базиса преобразования. Мгновенные частоты вычисляются от производных фазовых функций Гильбертовым преобразованием функций базиса [10].

Метод эмпирической модовой декомпозиции (EMD) предназначен для анализа нестационарных и нелинейных процессов. В отличие от Фурье и вейвлет-анализа EMD является интуитивным, прямым, адаптивным с апостериорно определяемым базисом, построенным по методу декомпозиции и зависящим от данных сигнала.

Декомпозиция основана на предположении, что любые данные состоят из различных простых внутренних модовых колебаний. Каждая внутренняя мода, линейная или нелинейная, представляет простое колебание, содержащее то же количество экстремумов и нулевых пересечений. Более того, колебания симметричны относительно локального среднего значения. В любой момент времени может сосуществовать множество внутренних колебаний, накладываемых друг на друга. Сами данные сигнала представляют собой сумму всех модовых колебаний [9].

Каждое из модовых колебаний представляет собой внутреннюю модовую функцию (IMF), определяемую правилами:

1) количество экстремумов функции (максимумов и минимумов) и количество нулей не должны отличаться более чем на единицу.
2) в любой точке среднее значение огибающей, построенной по локальным максимумам, и огибающей, построенной по минимумам, равно нулю.

Сравнительные характеристики методов Фурье, вейвлет и HHT анализа приведены в таблице 2.

4. Применение HHT к задаче сегментации речи

Так как речевой сигнал представлен нестационарным нелинейным процессом, к нему применимо преобразование Гильберта-Хуанга. Для проблемы сегментации основной задачей является поиск межфонемных переходов.

Как и в случае с вейвлет или Фурье анализа, результатом преобразования является разложение по ортогональному функциональному базису, но, в отличие от традиционных подходов, не априорному, а адаптивному.

Таблица 2. Сравнительные характеристики методов Фурье, вейвлет и HHT анализа

	Фурье	**Вейвлет**	**Гильберт-Хуанг**
Базис	Априорный	Априорный	Адаптивный
Частота	Свёртка: глобальные, неявные	Свёртка: локализованные, неявные	Дифференцирование: явные
Представление	Энергия-частота	Энергия-время-частота	Энергия-время-частота
Нелинейность	Нет	Нет	Да
Нестационарность	Нет	Да	Да
Выявление деталей	Нет	Дискретное: нет; непрерывное: да	Да
Теоретический базис	Полностью теоретический	Полностью теоретический	Эмпирический

Каждая базисная функция в данном случае соответствует внутреннему модовому колебанию. Соответственно на границах фонем можно ожидать быстрое изменение поведения IMF, т.е. концентрацию нулей производной сразу по нескольким модам. Возможна адаптация алгоритма [11] к HHT.

Представим общий алгоритм сегментации.

1. В качестве предобработки сигнал нормализуется: все отсчёты делятся на максимальное значение для установки единых пороговых значений для любых входных сигналов. Входной сигнал разбивается на фреймы по 32 мс, что соответствует 512 отсчётам при частоте дискретизации 16 кГц с перекрытием в половину окна.
2. На каждый фрейм накладывается оконная функция Хамминга для устранения дефектов на краях.
3. К каждому обработанному фрейму применяется HHT. Используется разложение до 8-ой моды и вычисляются $M_1 \dots M_8$.
4. Для каждой моды вычисляется производная MD.
5. Критерии выбора границ фонем: x_0 — предположительная граница фонемы, если $MD_i(x_1) = 0$; $MD_i(x_2) = 0$; $x_1 - x_2 < \mathrm{dif}_{opt}$; $x_0 = x_1 - x_2$, где $i = 2 \dots 8$, dif_{opt} экспериментально определяемое минимальное расстояние между экстремумами.
6. Для объединения результатов расстановки границ между модами все индексы объединяются в один вектор. Чтобы избежать ложных границ, устанавливается минимальный интервал фонемы — 25 мсек. Все границы, расположенные на расстоянии менее минимального, объединяются в группы, верной границей назначается сегмент по центру группы.

5. Результаты

Для первого теста были выбраны простые речевые фрагменты — дифоны. Алгоритм проверен на 20 различных дифонах, включающих сочетания как го-

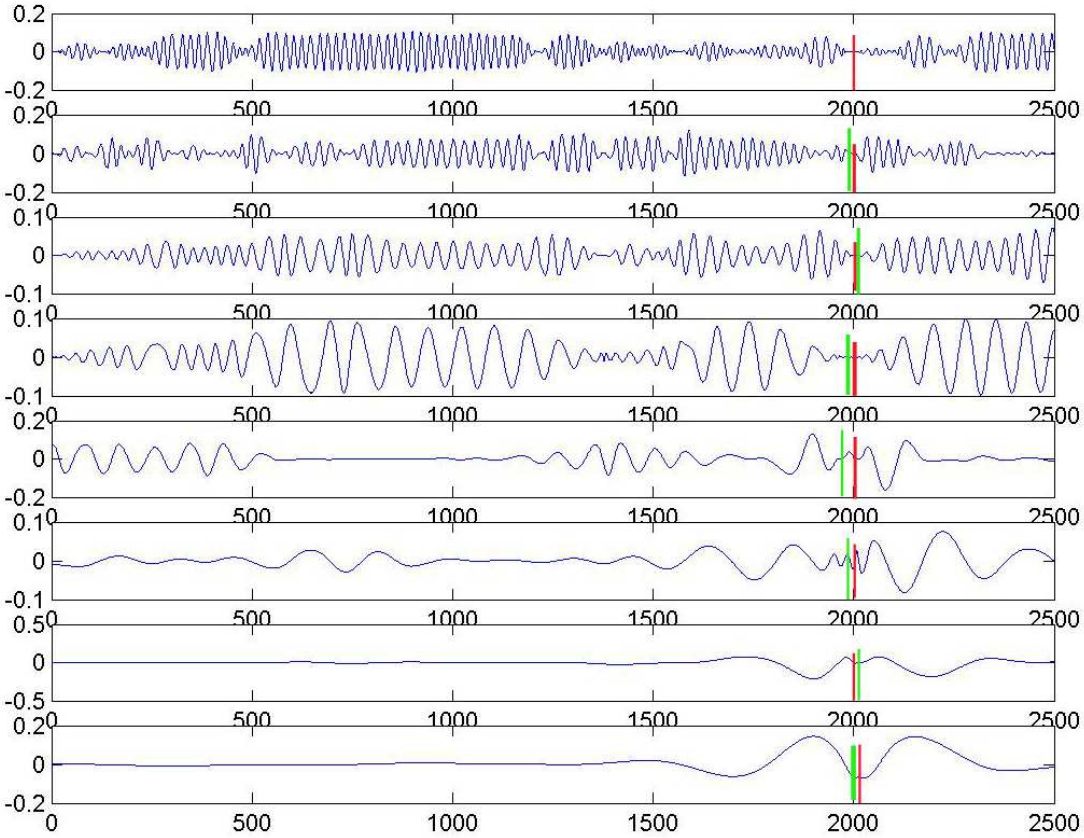

Рис. 2. Работа алгоритма на дифоне «а-о»

лосовых, так и неголосовых фонем. На рисунке 2 представлен пример работы алгоритма на дифоне «а-о».

Результаты ручной разметки соответствуют тёмной[1], а результаты разметки алгоритма — светлой[2] линии. По итогам тестирования следует отметить, что хорошо различима граница между гласными звуками, хотя эта проблема была одной из основных в алгоритме [11], напротив, границы между шипящими и гласными не всегда определяются корректно.

Заключение

Был предложен новый подход к задаче сегментации речевого сигнала, основанный на методе модовой декомпозиции. Перспективность применения HHT обусловлена природой речевого сигнала — его нелинейностью и нестационарностью. В качестве основного параметра определения точной границы сегмента используется скорость изменения поведения моды при последующем объеди-

[1]В цветной online-версии журнала соответствует красной линии.
[2]В цветной online-версии журнала — зелёная линия.

нении результатов расстановки границ между модами. Результаты показали правомерность применения данного подхода.

Для оптимизации алгоритма планируется использовать метод линейного предсказания для поиска оптимального условия для определения момента изменения характеристик мод и объединению результатов между модами.

Литература

1. Аграновский А.В., Леднов Д.А., Телеснин Б.А. Сегментация речи (математическая модель) // Информационные технологии. 1998. № 9. С. 24-28
2. Сорокин В.Н., Цыплихин А.И. Сегментация и распознавание гласных. // Информационные процессы. 2004. Т. 4. № 2. С. 202-220.
3. Силяков В.А., Красюк В.Н. Системы авиационной радиосвязи: Учебное пособие. СПб. : ГУАП, 2004. 160 с.
4. Авиационные радиосвязные устройства / Под ред. В.И. Тихонова. М. : ВВИА им. Н.Е. Жуковского, 1986. 442 с.
5. Перервенко Ю.С., Старченко И.Б. Анализ нелинейностей речевого сигнала // Радиоэлектроника и молодежь в XXI веке: материалы 11-го международного молодежного форума. Харьков : Изд-во ХНУРЭ, 2007. Ч.1. С. 289.
6. Фланаган Дж.Л. Анализ, синтез и восприятие речи / пер. с англ. А.А. Пирогова. М. : Связь, 1968. 397 с.
7. Teager H.M., Teager S.M. Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract // Speech Production and Speech Modelling, W.J. Hardcastle and A. Marchal, Eds., NATO Advanced Study Institute Series D, V.55, Bonas, France, July 1989.
8. Faúndez-Zanuy M., Kubin G., Kleijn W.B., Maragos P., McLaughlin S., Esposito A., Hussain A., Schoentgen J. Nonlinear speech processing: overview and applications // Engineering of Intelligent Systems. 2006. С. 1-5.
9. The Hilbert-Huang Transform and Its Applications / Eds By Norden E. Huang, Samuel S. Shen. Publisher: World Scientific Publishing Company. 2005. 324 p.
10. Давыдов А.В. Цифровая обработка сигналов: Тематические лекции. Екатеринбург : УГГУ, ИГиГ, ГИН. Фонд электронных документов, 2005.
11. Вишнякова О.А., Лавров Д.Н. Автоматическая сегментация речевого сигнала на базе дискретного вейвлет-преобразования // Математические структуры и моделирование. 2011. Вып. 23. С. 43-48.