ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 684.3

СПЕКТРАЛЬНЫЙ АНАЛИЗ И СОВРЕМЕННЫЕ РЕЧЕВЫЕ ТЕХНОЛОГИИ

канд. техн. наук, доц. В.И. СОЛОВЬЕВ
(Восточноукраинский национальный университет имени В. Даля, Луганск);
д-р техн. наук, проф. О.В. РЫБАЛЬСКИЙ
(Национальная академия внутренних дел, Киев);
д-р техн. наук, проф. В.К. ЖЕЛЕЗНЯК
(Полоцкий государственный университет)

Рассматривается применение спектрального анализа в современных системах идентификации диктора по голосу. Показано, что системы спектрального анализа на основе преобразования Фурье не обеспечивают возможности решения задач идентификации диктора на фонемном уровне. Создание таких систем возможно при использовании вейвлет-анализа на основе вейвлета Морле.

Введение. Большинство современных систем решения задач идентификации диктора по характеристикам голоса, распознавания речи основано на спектральном анализе аудиоинформации, базирующемся на математическом аппарате преобразования Фурье. Это обусловлено несколькими факторами: с одной стороны, известными нейрофизиологическими закономерностями обработки звуковой информации первичными слуховыми рецепторами; с другой – отсутствием более эффективных методов анализа и в какой-то степени историческими традициями в этой области.

Сегодня известны весьма эффективные системы распознавания характеристик голоса, разработки систем распознавания и текстового ввода речевой информации, однако достаточной ясности в принципиальных теоретических и практических вопросах речевых технологий нет. Среди них, с нашей точки зрения, наиболее важный – в чем заключается сущность эффективного декодирования речевой информации слуховым аппаратом на фонемическом уровне? Эта информация, безусловно, содержится в характеристиках звуковой волны, генерируемой речевым трактом. Но до настоящего времени отсутствует общепризнанная теория, позволяющая однозначно пояснить и смоделировать декодирование произвольной речевой информации на фонемическом уровне по характеристикам звуковой волны. Несмотря на большое число исследований и применение в последние двадцать лет мощной компьютерной техники, принципиальный прорыв в области физико-математических концепций эффективной обработки речевой информации не достигнут.

По мнению специалистов, во многом это обусловлено отсутствием эффективного математического инструмента для анализа речевой аудиоинформации.

Ниже излагается концепция, реализованная в системе инструментальной идентификации характеристик голоса диктора. Эта концепция базируется на представлении фрагментов речи в виде набора мультифрактальных структур. Для определения параметров мультифрактальных структур используется вейвлет-анализ со специальным базисом в виде двухпараметрического вейвлета Морле.

Постановка задачи исследования и разработки. Рассмотрим фрагмент речевого аудиофайла, показанного на рисунке 1 (фрагмент фонемы **[а]**).

Со времен Гельмгольца [1] известен очевидный факт, неоднократно отмечавшийся исследователями (в частности, классическая работа Фанта [2]), что большинство фонемических структур речи могут быть построены на основе близких геометрических компонент («атомарных» структур) звуковой волны. Эти структуры, как правило, ограничены временными интервалами, эквивалентными частоте основного тона речи. Геометрическое подобие этих структур является приближенным, но в большинстве случаев визуально очевидным (см. рис. 1). Временные интервалы, занимаемые этими структурами, обратны величине частоты основного тона и расположены в диапазоне от 2 до 15 мс. Данные атомарные структуры, рассматриваемые изолированно, не воспринимаются на слух ввиду их малой длительности звучания.

Следуя работам Мандельброта, эти структуры можно трактовать как атомарные составляющие мультифрактала [3–6]. При условии создания математической модели, обеспечивающей эффективное определение параметров «атомарных» структур и мультифрактала в целом, можно ожидать корректное описание и решение всех основных задач идентификации голоса и распознавания речи на фонемическом уровне.

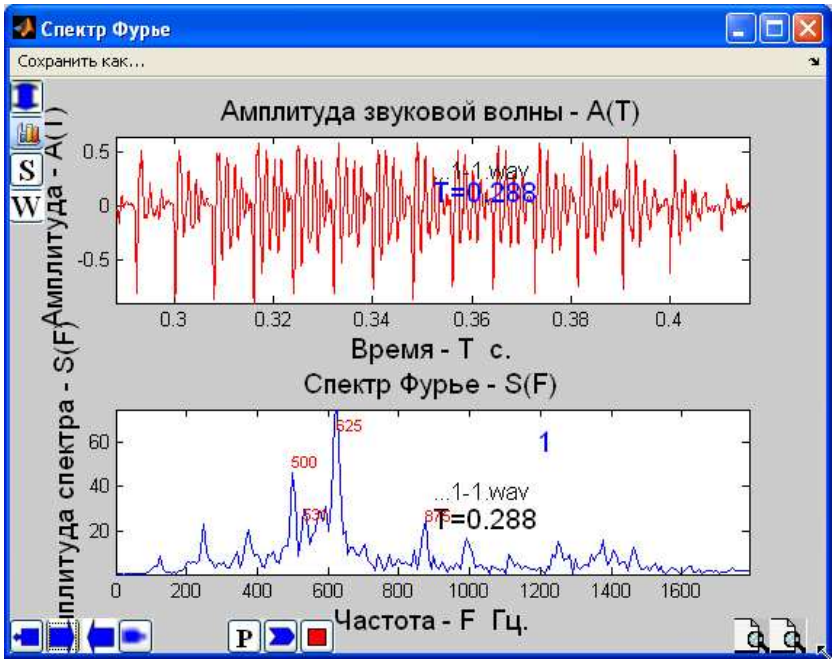

Рис. 1. Фрагмент фонемы **[a]**

До настоящего времени фактически единственным методологическим направлением построения математических моделей расчета параметров фонем является применение той или иной модификации спектрального анализа на основе преобразования Фурье. Однако применение этого метода спектрального анализа имеет ряд серьезных изъянов (известных со времен Гельмгольца). Спектр на достаточно больших интервалах времени (см. рис. 1) часто представляет модифицированную картину энергетической доли частот в силу эффекта расщепления частот при амплитудных модуляциях (или биениях).

Как известно, спектр сигнала «аппаратных» преобразователей спектра слухового аппарата – волосковых клеток – существенно отличается от спектра Фурье [7–14], в том числе (что весьма важно) по расположению формант в частотной области.

Вторым важным фактором неудобства спектрального анализа Фурье является фиксированный шаг по частоте спектра при малых временных окнах анализа. Так, для наиболее интересных интервалов в диапазоне от 2 до 20 мс такой анализ требует шага по частоте 50–500 Гц соответственно. Это свойство в первую очередь обусловлено применением ортогонального быстрого преобразования Фурье. Слуховой аппарат человека различает частоты с шагом менее 1 Гц (в диапазоне частот до 500 Гц).

Разумеется, что физико-математические модели преобразования речевой информации не обязательно должны строго следовать нейрофизиологическим закономерностям. Однако анализ речевой информации на основе преобразования Фурье на фонемическом уровне указывает на весьма высокую вариабельность спектров для одних и тех же фонемических структур, что является главным препятствием на пути эффективной оценки параметров фонем. Следовательно, решение задачи математического описания «атомарных» структур фонем как мультифракталов выступает определяющей предпосылкой решения задач инструментальной идентификации диктора по характеристикам голоса и распознавания фонем речи по характеристикам звуковой волны.

Выявление атомарных составляющих фонем как мультифракталов

Оценка параметров атомарной структуры. Известно, что альтернативой спектральному анализу Фурье является вейвлет-анализ [15]. Однако многочисленные попытки применения методов вейвлет-анализа для обработки речевой информации до настоящего времени не дали серьезных результатов.

Препятствием на этом пути являются, с одной стороны, сложность физической интерпретации результатов исследований для большинства вейвлетных базисов; с другой – применение вейвлет-базисов, имеющих близкое сродство с речевыми фрагментами (например, базис Морле), затруднено ввиду высокой вычислительной сложности.

Рассмотрим фрагменты речи в аудиоданных как дискретный временной ряд амплитуды звуковой волны. Поставим задачу выделения характеристик самоподобных структур в полученном временном ряду для выделенных в предыдущем разделе атомарных фрагментов речи.

Для выявления подобных структур используем методы вейвлет-анализа [15]. С этой целью выберем комплексный вейвлет Морле [15]:

$$C_{mor}(t_i, T_k, F_b, F_c) = (\pi F_b)^{0,5} \exp(2j\pi F_c t_i) \exp(-(t_i - T_k)^2 / F_b), \quad (1)$$

где t_i – дискретные временные отсчеты; T_k – временной отсчет, соответствующий центральной части временного окна, – комплексная единица; F_b – параметр ширины вейвлета; F_c – центральная частота вейвлета;

Пусть $A(t_i)$ – значение амплитуды звуковой волны фрагмента речи аудиофайла в момент времени t_i.

Рассмотрим временное окно фрагмента речи с интервалом δT менее 20 мс. Параметр ширины комплексного вейвлета Морле F_b выбран на основе экспериментальных исследований постоянным для всех преобразований. Его величина выбиралась из условия практического затухания абсолютных значений вейвлета Морле при значениях $t_i - T_k$, равных $\delta T/2$. Будем для каждого фрагмента речи вычислять свертку вейвлета Морле с фрагментом временного ряда амплитуды звуковой волны в виде

$$C(T_k, F_b, F_c) = (1/N)_{abs} \left(\sum_{t_{ij}=0}^{N_m} C_{mor}(t_i, T_k, F_b, F_c) A(t_i) \right), \quad (2)$$

где $C(T_k, F_b, F_c)$ – значение модуля коэффициента вейвлет-преобразования; N – количество дискретных отсчетов на интервале δT временного окна.

При фиксированном параметре ширины F_b комплексного вейвлета Морле значение модуля является функцией частоты F_c вейвлета Морле и положения временного окна по времени T_k. Типичный график пространственной скейлограммы $C(T_k, F_b, F_c)$ в функции F_c и T_k для рассматриваемых нами фрагментов речи представлен на рисунке 2.

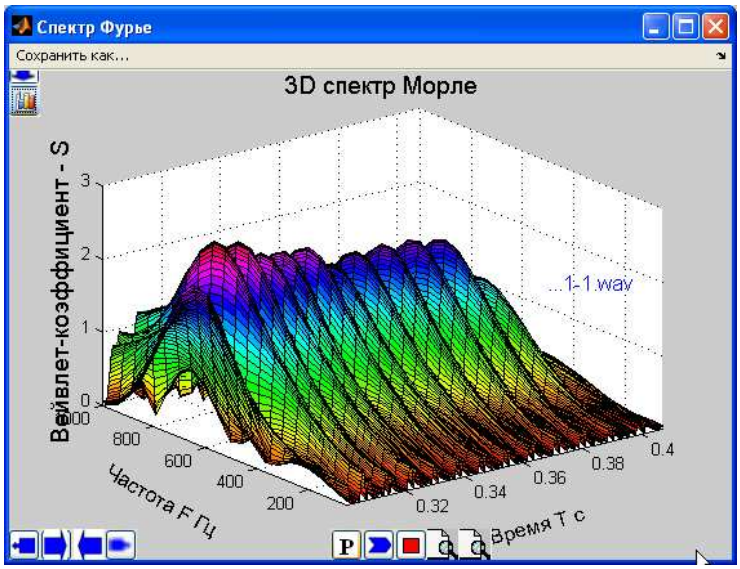

Рис. 2. Скейлограмма фрагмента речи

Данный график представляет собой трехмерный спектр Морле для фрагмента фонемы **[а]**, показанной на рисунке 1.

Частотно-временное представление фрагмента речи в виде пространственной скейлограммы на основе базиса Морле обладает рядом важных особенностей, позволяющих существенно повысить эффективность выявления самоподобных структур. В частности, локальные максимумы вейвлет-преобразования характеризуются как весьма информативные для анализа атомарных составляющих мультифракталов в аудиоинформации.

Модель выявления самоподобных структур «атомарных» фрагментов фонем. Анализ скейлограмм показывает, что расположение «хребтов» скейлограмм по временному параметру на рисунке 2

строго соответствует локальным экстремумам амплитуды звуковой волны во временной области. Причем эти локальные экстремумы соответствуют всплескам амплитуды звуковой волны, обусловленным частотой основного тона. Но наиболее существенной особенностью характеристик «хребтов» является форма «хребта» на его «вершине». Исследования показывают, что после соответствующей нормировки функции $C(T_k, F_b, F_c)$ при фиксированных T_k и F_b на вершине «хребта» эти функции обладают высокой степенью геометрического подобия. При этом форма нормированных «хребтов» индивидуальна, что обусловлено характеристиками голоса, и полностью отображает атомарные составляющие фонемы как мультифрактала.

Важным фактором высокого сродства базиса Морле с самоподобными структурами в речевых фрагментах является более высокая степень гладкости скейлограммы по сравнению, например, с аналогичным преобразованием Фурье.

Так, на рисунках 3 и 4 представлены двумерные срезы трехмерной скейлограммы (для фрагмента речи, представленного на рисунке 1) в координатах времени и частоты. Более высокая степень гладкости функций обеспечивает весьма эффективную возможность математического анализа параметров «хребтов».

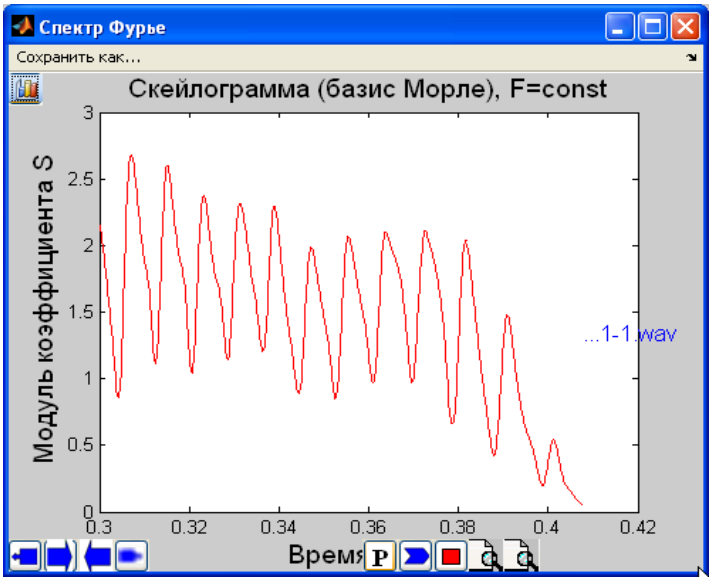

Рис. 3. Двумерная скейлограмма по времени

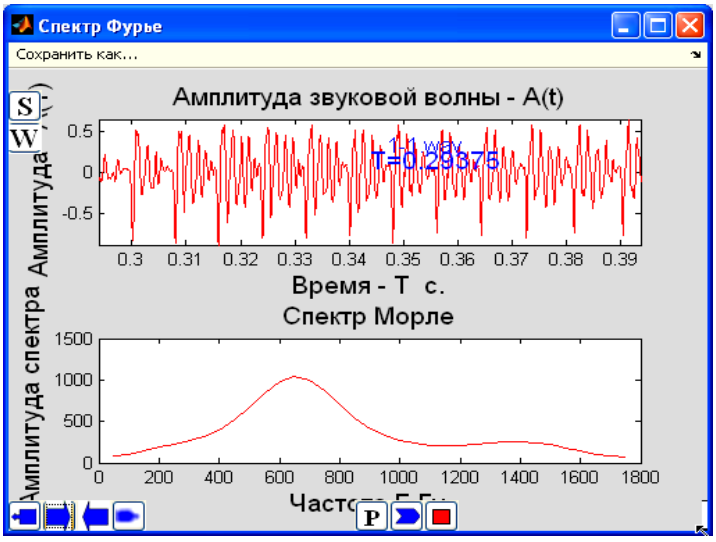

Рис. 4. Двумерная скейлограмма по частоте

Как видно из рисунка 4, двумерный спектр Морле в разрабатываемой модели принципиально отличается от спектра Фурье на рисунке 1. Как правило, при формантном подходе на фонемах в спектре

Морле возможно выявить не более трех-четырех формант. При этом максимумы формант по частоте существенно могут отличаться от их максимумов в спектре Фурье. Такое отличие объясняется принципиальной разницей подхода на основе вейвлет-анализа и спектрального преобразования Фурье. В разработанном подходе практически отсутствует влияние амплитудных модуляций на частоты.

Заключение. В результате проведенного исследования разработана модель выявления самоподобных структур в речевых аудиофайлах на основе вейвлетного базиса Морле. Модель позволяет реализовать эффективную идентификацию «атомарных» составляющих фрагментов речи в частотно-временной области.

ЛИТЕРАТУРА

1. Von Helmholtz, H. Die Lehe von Tonempfindungen / H. von Helmholtz. – Brannschweig, Vieweg, 1863.
2. Фант, Г. Анализ и синтез речи / Г. Фант; пер. с англ. В.С. Лозовского и Н.В. Бахмутовой под ред. Н.Г. Загоруйко. – Новосибирск: Наука, Сиб. отд., 1970. – 167 с.
3. Mandelbrot, B. Statistical Methodology for Non-Periodic Cycles: From the Covariance to R/S Analysis / B. Mandelbrot // Annals of Economic Social Measurement. – 1972. – № 1.
4. Mandelbrot, B. The Fractal Geometry of Nature / B. Mandelbrot. – New York: W. H. Freeman, 1982.
5. Mandelbrot, B. A Multifractal Walk Down Wall Street / B. Mandelbrot // Scientific American, 1999.
6. Mandelbrot, B. Robustness of the rescaled range R/S in the measurement of non-cycling long-run statistical dependence / B. Mandelbrot // Water Resources Research. – 1969. – № 5. – P. 967–988.
7. Психоакустические аспекты восприятия речи. Механизмы деятельности мозга / под ред. Н.П. Бехтеревой. – М.: Наука, 1988. – 504 с.
8. Цвикер, Э. Ухо как приемник информации / Э. Цвикер, Р. Фельдкеллер; пер. с нем. под ред. Б.Г. Белкина. – М.: Связь, 1971. – 225 с.
9. Алдошина, И.А. Основы психоакустики / И.А. Алдошина // Звукорежиссер. – 2000. – № 6. – С. 36–40.
10. Сорокин, В.Н. Теория речеобразования / В.Н. Сорокин. – М.: Радио и связь, 1985. – 312 с.
11. Цвикер, Э. Ухо как приемник информации / Э. Цвикер, Р. Фельдкеллер; пер. с нем. под ред. Б.Г. Белкина. – М.: Связь, 1971. – 225 с.
12. Phase in speech pictures / A.V. Oppenheim [et al.] // in Proc. IEEE Int. Conf Acoust. Sprech and Signal Processing. – 1979. – Apr. – P. 632–637.
13. Аграновский, А.В. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов / А.В. Аграновский, Д.А. Леднов. – М.: Радио и связь, 2004. – 434 с.
14. Фланаган, Дж. Анализ, синтез и восприятие речи / Дж. Фланаган; пер. с англ. под ред. А.А. Пирогова. – М.: Связь, 1968. – 396 с.
15. Мала, С. Вейвлеты в обработке сигналов / С. Мала. – М.: Мир, 2005. – 670 с.

Поступила 16.12.2013

SPECTRAL ANALYSIS AND ADVANCED SPEECH TECHNOLOGIES

V. SOLOVYOV, O. RYBALSKY, V. ZHELEZNYAK

Application of spectrology is considered in the modern systems of authentication of announcer on voice. It is shown that the systems of spectrology on the basis of transformation of Fourier do not provide possibility of decision of tasks of authentication of announcer at phoneme level. It is shown that creation of such systems is possible at the use of wevlet-analysis on the basis of wevlet Morlet.