УДК 621.391

Е.Г. ЖИЛЯКОВ, д-р техн. наук, БелГУ (г. Белгород, Россия),
С.П. БЕЛОВ, канд. техн. наук, БелГУ (г. Белгород, Россия),
Е.И. ПРОХОРЕНКО

О СЖАТИИ РЕЧЕВЫХ СИГНАЛОВ

У статті запропоновано один з шляхів рішення проблеми зменшення об'єму при бітовому представленні звукових даних. Це – комплексний метод, який досягнуто завдяки застосуванню основаному при видвіженні на перший план та кодуванні сигналів звуку та пауз за допомогою моделей авторегресії з наступним сжиманням сигналу за допомогою субполос кодування.

In article is offered one of the ways of decision a problem of reduction of volume of bit presentation voice data. This is the complex method reached by using, founded on highlighting and coding in input an voice a signal a pause with the help of models an autoregression with the following compression of signal by means of sub-band of coding.

Постановка проблемы. Одна из главных целей исследований свойств речевых сигналов – определение возможности уменьшения объемов кодированных речевых данных без ухудшения качества воспроизведения речи. Данная задача представляет интерес как в контексте уменьшения скорости передачи для данной ширины полосы передачи в системах телекоммуникаций, так и для записи звуковых файлов, содержащих речевые данные, на жесткие носители информации.

Анализ литературы. Речевые сигналы представляются посредством речевых данных, под которыми, в контексте данной работы, понимается результат кодирования звуков речи с помощью некоторых методов. В современных цифровых системах речевые сигналы представляют в дискретном виде и сохраняют, в соответствии с некоторыми правилами, значения дискретных отсчетов. В процессе записи исходные речевые данные должны быть преобразованы таким образом, чтобы сократить объем их битового представления.

При записи речевых данных в файл сокращение объема может быть достигнуто за счет кодирования пауз, объем которых в речевых данных значителен [1]. Кодирование пауз заключается в определении интервала, на котором отсутствуют звуки речи, фиксации начала этого интервала и его длительности. Кроме того, для воспроизведения речи с комфортным звучанием необходимо определить некоторые параметры этого интервала, например значения математического ожидания и наименьшего среднеквадратичного отклонения. Очевидно, что длительность пауз между словами, фразами зависит от дикторов. Кроме того, известно, что речь состоит из фонем перемежающихся паузами, которые так же целесообразно кодировать. Например, в системах IP-телефонии и мобильной связи применяются кодеры с переменной скоростью кодирования речевого сигнала, в основе которых лежит классификатор входного сигнала, который выделяет во входном речевом сигнале активную речь и паузы, определяющий степень его информативности и, таким образом, задающий метод кодирования и скорость передачи речевых данных. Наиболее простым классификатором речевого сигнала является детектор активности речи (VAD – Voice Activity Detector), который выделяет во входном речевом сигнале активную речь и паузы. Фрагменты, классифицированные как паузы, кодируются и передаются с очень низкой скоростью (порядка 0,1 – 0,2 кбит/с) или не передаются вообще. В некоторых случаях целесообразно более детально осуществлять классификацию фрагментов, соответствующих активной речи [2, 3].

Другим методом сокращения объема речевых данных может служить субполосное кодирование, с помощью которого осуществляется сжатие речевого сигнала за счет учета его частотных свойств. То есть при неравномерном заполнении частотного диапазона за счет разделения сигнала на субполосы возможно уменьшить суммарное число бит, отводимых на кодирование сигнала. При этом погрешность в смысле дисперсии ошибки восстановления не должна быть большой [4, 5].

Цель статьи. В данной работе рассматривается проблема сжатия речевых данных. При этом основное внимание уделяется задаче обнаружения и кодирования пауз. Кроме того, на уровне экспериментальных исследований оценивается возможность сжатия данных с исключением пауз на основе субполосного кодирования [6].

1. Основы метода обнаружения пауз

Пауза в речевом сигнале – отрезок, содержащий более или менее нерегулярные случайные изменения, т.е. процесс, который можно отнести к классу случайных. Если принять, что вероятностная структура паузы не изменяется со временем, то тогда сигнал можно считать случайным стационарным процессом.

Сигнал, соответствующий звукам, не является стационарным, так как формируется при активном воздействии речевого аппарата. Процедура обнаружения пауз может быть основана на принципе обнаружения отличий характеристик сигналов на данном интервале по сравнению с характеристиками сигнала в паузе.

Для описания характеристик сигнала в паузе, в данной работе используется модель авторегрессии [7], или стохастическое разностное уравнение, которое можно представить следующим образом:

$$y(n) - m = \sum_{k=1}^{p} \alpha_k (y(n-k) - m) + \sigma_0 u_t , \qquad (1)$$

где m – математическое ожидание $y(n)$; u_t – некоррелированная последовательность с единичной дисперсией и нулевым математическим ожиданием; σ_0 – параметр, определяющий уровень среднеквадратичной погрешности предсказания на основе линейной комбинации вида:

$$y(n) = \sum_{k=1}^{p} \alpha_k (y(n-k) - m) . \qquad (2)$$

Отрезки речевого сигнала, которые не относятся к паузе, будут отличаться по структуре от сигнала в паузе, т.е. для каждого звука адекватна своя модель генерации, возможно отличная от модели авторегрессии. Таким образом, можно сформулировать следующие гипотезы:

H_0: отрезок сигнала генерируется по схеме (1) с известными параметрами α_k, m, σ_0, p.

H_1: отрезок сигнала генерируется по схеме, отличной от (1).

Для проверки этих гипотез необходимо ввести решающую функцию (РФ). В данной работе были использованы результаты, полученные в [8], где найдена, не имеющая мертвых зон векторная решающая функция $S_m(\tau)$, в качестве компонент которой используются статистики максимальной чувствительности:

$$S_1(\tau) = \sum_{k=1}^{\tau} z(k) ; \qquad (3)$$

$$S_2(\tau) = \frac{1}{2} \sum_{k=1}^{\tau} z^2(k) - \frac{\tau}{2} ; \qquad (4)$$

$$S_i(\tau) = \sum_{k=1}^{\tau+2-i} z(k) z(k+i-2), \quad i = 3, ..., m , \qquad (5)$$

где $z(i)$ последовательность вида

$$z(i) = \sum_{k=0}^{p} \alpha_k (y(i-k) - m) / \sigma_0 . \qquad (6)$$

Соотношения для границ доверительных интервалов РФ имеют вид

$$a_i(\tau) = -b_i(\tau), \qquad b_i(\tau) = k\sqrt{\tau - i + 2} , \qquad (7)$$

где k – некоторый коэффициент.

Реально модель авторегрессии и ее параметры неизвестны, и их можно только оценить по значениям реализаций случайных последовательностей. В данной работе рассматривается один из возможных подходов к решению этой задачи с применением принципа адаптации к реализациям пауз речевых сигналов, используемых при формировании звуковых файлов. Для оценки параметров модели (1) используется метод наименьших квадратов [9].

2. О субполосном преобразовании речевых сигналов

Пусть далее $\vec{x} = (x_1, ..., x_N)^T$ – вектор, компоненты которого представляют собой вещественнозначные отсчеты речевого сигнала на некотором интервале регистрации так, что индексы однозначно связаны с моментами эквидистантной дискретизации.

Соответствующими большими буквами будем обозначать трансформант Фурье

$$X(v) = \sum_{k=1}^{N} x_k e^{-j(k-1)v} , \qquad (8)$$

где v – нормированная круговая частота,

$$v = \omega\Delta = 2\pi f\Delta, \qquad (9)$$

где f – частота в герцах, а Δ – длина интервала временной дискретизации речевого сигнала, которая предполагается эквидистантной.

Известно [2], что модуль спектра речевого сигнала обладает рядом особенностей, среди которых отметим:

– наличие максимумов, которые соответствуют почти периодическим компонентам (основной тон и форманты);

– нестационарность (изменчивость), которая обусловлена различиями генерации сигнала в зависимости от произнесенного звука;

– хорошей повторяемостью формы в случае произнесения одного и того же звука.

Поэтому для анализа речевых сигналов достаточно широкое распространение получил подход на основе так называемого кратковременного спектра, когда вычисляется последовательность квадратов модулей вида

$$\left|\mathrm{X}_{r,k}(v_i)\right|^2 = \left|\sum_{k=1}^{M+r} x_k e^{-j(k-r)v_i}\right|^2, \qquad (10)$$

где v_i – конкретное значение частоты из некоторого множества, например

$$v_i = (i-1)\delta, \quad \delta = \frac{\pi}{L};\ r = 1, 2, \ldots, N - M. \qquad (11)$$

Здесь M – длина интервала анализа. При выборе M ищется компромисс между стремлением, с одной стороны, иметь возможность отделить проявления компонент с различными квазипериодами (частотное разрешение, которое при неизменности характеристик улучшается с ростом M), а, с другой стороны, учитывается существенная нестационарность сигнала, связанная с неизбежной сменой произносимых звуков. Заметим, что артикуляция гортани меняется и при генерации одного и того же звука.

На основе вычисления (10) строятся двумерные графики, ось абсцисс которых представляет собой временные отсчеты (r), а ось ординат – отсчеты частоты v_i.

Эти графики принято называть спектрограммами. В настоящее время процедура их создания реализована в виде программной поддержки, что свидетельствует о востребованности такого анализа.

Отметим, однако, что более целесообразно вычислять значения интегралов вида:

$$P_{r,M,i} = \int_{v \in V_i} \left|\mathrm{X}_{r,M}(v)\right|^2 dv; \qquad (12)$$

$$V_i = \bigcup \left[v_{2i} - v_{1i}, v_{2i}, v_{2r} > 0, v_{1r} \geq 0\right]. \qquad (13)$$

Имеются в виду доли энергии анализируемого отрезка сигнала, которые сосредоточены в частотных интервалах вида (13).

Нетрудно доказать [10] справедливость представления

$$P_{r,M,i} = \vec{y}_{r,M}^{\,T} A_i \vec{y}_{r,M}, \qquad (14)$$

где $\vec{y}_{r,M} = (x_r, \ldots, x_{r+M-1})^T$; $A_i = a_{m,n}^i$, $m, n = 1, \ldots, M$;

$$a_{n,m}^i = \begin{cases} \dfrac{\sin v_{2i}(m-n) - \sin v_{1i}(m-n)}{\pi(m-n)}, & m \neq n, \\ \dfrac{v_{2i} - v_{1i}}{\pi}, & m = n. \end{cases} \qquad (15)$$

Отметим, что представление (14) позволяет вычислять значения интеграла вида (12) в области трансформант Фурье.

Таким образом, соотношение (14) представляет собой новый инструмент, позволяющий анализировать спектральные плотности отрезков речевых сигналов.

Очевидно, что соотношение (14) позволяет модифицировать спектрограммы, которые будут более адекватны процессу слухового восприятия звука.

С позиций сжатия битовых представлений отрезков сигналов существенной является возможность пренебречь частью частотных составляющих, энергии которых малы (в относительном смысле). При этом важно осуществить переход от векторов отсчетов во временной области $\vec{y}_{r,M}$ к векторам $y\vec{y}_{r,M}$, которые непосредственно отражают свойства в различных интервалах оси частот (субполосное кодирование).

Положим:

$$\vec{z}_{r,M,i} = B_i \vec{y}_{r,M}, \qquad (16)$$

где $B_i = b^i_{n,m}$, $m = 1,...,M;\ n = 1,...,G_i$;

$$G_i = \frac{M(v_{2i} - v_{1i})}{\pi}; \qquad (17)$$

$$b^i_{n,m} = \begin{cases} \dfrac{G_i}{\pi} \dfrac{(-1)^n \sin(v_{2i} \cdot n) + \sin(v_{1i} \cdot n)}{nG_i - m}, & m \neq nG_i, \\ \dfrac{G_i}{\pi}(v_{2i} - v_{1i}) \cos(nG_i v_{1i}), & m = nG_i. \end{cases}$$

Можно доказать справедливость утверждения

$$\int_{v \in V_i} \left| Y_{r,M}(v) - Z_{r,M}(G_i v) \right|^2 dv = \min \int_{v \in V_i} \left| Y_{r,M}(v) - W(G_i v) \right|^2 dv, \qquad (18)$$

где $W(G_i v)$ – спектр любого другого вектора той же, что и $\vec{z}_{r,M,i}$ размерности.

Иными словами, спектр вектора $\vec{z}_{r,M,i}$ является наилучшим приближением в смысле евклидовой нормы разностей для спектра исходного вектора $\vec{y}_{r,M}$ в выбранном частотном интервале и в этом смысле вектор вида (16) является оптимальным.

Если теперь выбрать разбиение оси частот на L непересекающихся интервалов вида

$$V_{ii} = 0, \quad v_{2,i} = v_{1,i+1},\ i = 1, 2, \ldots, L, \qquad (19)$$

то вектор

$$zz_{r,M} = BB \cdot \vec{y}_{r,M} = \left(\vec{z}^T_{r,M,1}, ..., \vec{z}^T_{r,M,L} \right)^T, \qquad (20)$$

где BB – блочная матрица размерности M x M ,

$$BB = \begin{pmatrix} B_1 \\ B_2 \\ \ldots \\ B_L \end{pmatrix}, \qquad (21)$$

будет оптимальным в смысле возможности наилучшей аппроксимации исходного спектра в каждом из частотных интервалов (в смысле (18)).

Соотношение (21) естественно называть субполосным преобразованием.

Важно, что в виду неособенности матрицы BB имеет место обратная операция, то есть:

$$\vec{y}_{r,M} = BB^{-1} zz_{r,M} . \qquad (22)$$

Иными словами, соотношения (20) и (22) представляют собой прямое и обратное субполосное преобразование.

Если теперь, исходя их некоторых соображений (например, энергетических), в векторе $\vec{zz}_{r,M}$ обнулить часть компонент, то это будет способствовать сжатию данных, тогда как представление (22) позволит восстановить исходный вектор.

3. Экспериментальное исследование эффективности сжатия

Для оценки работоспособности предложенного метода определения в речевых сигналах отрезков, соответствующих паузам между словами, фразами и фонемами были проведены вычислительные эксперименты.

В виде файлов речевых данных была записана лекция, прочитанная двумя различными дикторами.

С использованием разработанных прототипов программных средств, позволяющих реализовывать предложенный выше метод кодирования пауз, в каждом файле речевых данных был выделен отрезок данных, достоверно соответствующий отсутствию речевого сигнала, и определены оценки параметров модели (1). С использованием этих оценок вычислялась решающая функция $S_2(\tau)$ (4).

Таблица

Объем, Мбайт			Качество воспроизведения после декодирования
Исходный файл	Файл с закодированными паузами	Файл, полученный в результате субполосного кодирования	
6,25	3,52	0,46	Неразборчиво
		0,76	Низкое
		1,3	Слова разборчивы, но приходится незначительно напрягать внимание
		1,9	Высокое
5,24	3,86	0,53	Неразборчиво
		0,85	Низкое
		1,38	Слова разборчивы, но приходится незначительно напрягать внимание
		2,14	Высокое

Выбор значений τ и k осуществлялся исходя из принципа максимальной чувствительности к отличию отрезка сигнала от паузы. Это позволяет, до определенной степени, гарантировать отсутствие искажений данных, относящихся к отрезкам, соответствующим звучанию голоса. Искажения возникают при ошибочном исключении таких данных. Вместе с тем, основной целью является исключение максимального объема данных, относящихся к паузам. Представляется, что этим требованиям будет отвечать РФ, значения которой в паузе близки к границам вида (7), но пересекают их с заданной вероятностью ложной тревоги, уровень которой, в свою очередь, гарантирует малую степень искажений речевых сообщений. В соответствии с выражением (7) качество процедуры обнаружения и удаления пауз будет зависеть от длины отрезка τ и параметра k. Экспериментально установлено, что при выборе τ порядка 3 и k порядка 30 обеспечивается приемлемая степень искажения речевых сообщений и регистрируется достаточный объем данных, соответствующий паузам.

Если значения РФ на i-м интервале не превышали пороговых значений (7), то значения дискретных отсчетов речевых сигналов на данном интервале удалялись, а пауза кодировалась значениями номеров начальных отсчетов и длительностью. Случаи отнесения к паузе отрезков, длина которых меньше некоторой, игнорировались, так как в противном случае при воспроизведении речи появляются шумовые эффекты в виде тресков. Таким образом, формировались последовательности, представляющие собой речевые данные, не содержащие отрезков, соответствующих паузам речевого сигнала.

Сжатие оставшихся битовых представлений сигналов производилось посредством субполосного кодирования с различным числом обнуляемых компонент, что оказывало существенное влияние на качество воспроизведения восстановленного сигнала. Качество воспроизведения речи, полученной после восстановления речевых данных, оценивалось несколькими независимыми экспертами. В таблице приведены количественные значения объемов файлов речевых данных, полученных на различных этапах сжатия. При этом можно обеспечить изменение коэффициента сжатия в широких пределах.

Выводы. Результаты вычислительного эксперимента свидетельствуют о том, что предлагаемый комплексный метод позволяет уменьшать объем файлов речевых данных от пяти до десяти раз при сохранении высокой степени качества воспроизведения речевых сообщений.

Список литературы: 1. *Росляков А.В., Самсонов М.Ю., Шибаева И.В.* IP-телефония. – М.: Эко-Тредз, 2001. – 250 с. **2.** *Шелухин О.И., Лукьянцев Н.Ф.* Цифровая обработка и передача речи. – М.: Радио и связь, 2000. – 456 с. **3.** *Артюшенко В.М., Шелухин О.И., Афонин М.Ю.* Цифровое сжатие видиоинформации и звука. – М.: Издательско-торговая корпорация "Дашков и К", 2003. – 426 с. **4.** *Сергиенко А.Б.* Цифровая обработка сигналов. – СПб.: Питер, 2005. – 752 с. **5.** *Смоленцев Н.К.* Основы теории вейвлетов. Вейвлеты в MatLab. – М.: ДМК Пресс, 2005. – 200 с. **6.** *Жиляков Е.Г., Попов И.Г., Чижов И.И.* Программный комплекс сжатия–восстановления звуковых файлов при передаче по каналам Интернет. – Свидетельство № 4314 об отраслевой регистрации разработки от 22 февраля 2005 года. – Федеральное агентство по образованию. **7.** *Андерсен Т.* Статистический анализ временных рядов. – М.: Мир, 1976. – 760 с. **8.** *Жиляков Е.Г., Шпилевский Э.К.* Статистики максимальной чувствительности в задаче обнаружения изменений параметров процессов авторегрессии // Заводская лаборатория. – 1992. – № 7 – С. 31 – 34. **9.** *Жиляков Е.Г., Корсунов Н.И., Лагода Д.П.* Методы и алгоритмы обработки экспериментальных данных в атомно-абсорбционной спектрометрии. – К.: Наукова думка, 1992. – 122 с. **10.** *Жиляков Е.Г., Попов И.Г., Чижов И.И.* О субполосном кодировании сигнала // Вестник НТУ "ХПИ", 2004. – № 46. – С. 10 – 19.

Поступила в редакцию 11.10.2005