УДК 621.391

А.В. УСТИНОВА, БелГУ (г. Белгород),
Д.В. УРСОЛ, БелГУ (г. Белгород)

О СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ НА ОСНОВЕ ЧАСТОТНЫХ ПРЕДСТАВЛЕНИЙ

Розглядаються нові алгоритми сегментації речових сигналів на відрізку пауза/звук, засновані на використанні нового методу обчислювання точних значень часток енергії відрізків сигналів у заданих частотних інтервалах.

The new algorithms of speech signals segmenting on a part pause/sound are considered, which are founded on use the of a new calculation method of proper values of energy parts of signals length of in given frequency intervals.

Постановка проблемы. Постоянно нарастающая интенсивность использования информационно-телекоммуникационных систем (ИТС) для речевого взаимодействия привела к необходимости разработки способов минимизации затрат на хранение и передачу речевых данных, что достигается за счёт уменьшения объёмов их битовых представлений.

Особенностью речевых сигналов является высокая доля пауз. Суммарная продолжительность перерывов в среднем занимает около 15 % от продолжительности слитной речи, а в режиме диалога 56 % от общей длительности. Кроме того, речь содержит множество кратких перерывов длительностью от 5 до 200 мс, существующих как внутри слов, так и между словами в слитной речи [1]. Поэтому удаление из файла блока данных, соответствующих паузам, позволяет существенно уменьшить объемы битовых представлений речевых сообщений. Также важно не исказить речь за счет удаления части собственно звуковых данных, что может являться следствием несовершенства применяемой информационной технологии удаления данных паузы.

Анализ литературы. Применяемые (в основном в телекоммуникациях) в настоящее время решающие процедуры обнаружения пауз основываются на использовании так называемых фильтров линейного предсказания [1, 2]. Среди несовершенств такого подхода можно выделить: принципиальную невозможность построения фильтра линейного предсказания конечного порядка для "белого" шума; наличие в решающей функции "мертвых зон", когда изменение одних параметров компенсируется изменениями других; возможное совпадение максимумов энергетических спектров шума и звука, что приводит к совместному их подавлению и ошибочному отнесению анализируемого участка к паузе и т.д. [3, 4].

Цель статьи. Основное отличие между сигналом, соответствующим паузе, и звуковыми данными заключается в распределении энергий по частотному диапазону. В данной работе рассматривается метод обнаружения

пауз, который адекватно отображает это отличие, что при прочих равных условиях создает предпосылки повышения достоверности принимаемых решений.

1. Оценка энергии речевого сигнала

В данной работе предлагается новый метод вычислений долей энергии отрезков речевых сигналов, соответствующих заданным частотным диапазонам. Основная суть метода заключается в следующем.

Пусть компоненты вектора

$$\vec{x} = (x_1, ..., x_N)^T \tag{1}$$

представляют собой значения некоторого сигнала (функции времени).

Положим далее

$$X(\upsilon) = \sum_{k=1}^{N} x_k e^{-j(k-1)\upsilon} , \tag{2}$$

т.е. $X(\upsilon)$ представляет собой трансформанту Фурье отрезка отсчетов сигнала (вектора), для частотного интервала

$$V = \left[-\upsilon_2, -\upsilon_1\right) \cup \left[\upsilon_1, \upsilon_2\right). \tag{3}$$

Тогда выражение

$$P_V(\vec{x}) = \frac{1}{2\pi} \int_{\upsilon \in V} \left|X(\upsilon)\right|^2 d\upsilon \tag{4}$$

представляет собой долю энергии отрезка сигнала (евклидовой нормы вектора), соответствующую частотному интервалу (3).

В работе [7] показано, что если в правую часть соотношения (2) подставить определение (4), то в результате преобразований получим

$$P_V(\vec{x}) = \vec{x}^T \mathbf{A}_V \vec{x} , \tag{5}$$

где $\mathbf{A}_V = \{a_{ik}\}$, $i = 1, \ldots, N$, $k = 1, \ldots, N$ – симметричная матрица, элементы которой определяются как

$$a_{ik} = \begin{cases} \dfrac{\sin\left[\upsilon_2 (i-k)\right] - \sin\left[\upsilon_1 (i-k)\right]}{\pi (i-k)}, & i \neq k, \\ \dfrac{\upsilon_2 - \upsilon_1}{\pi}, & i = k. \end{cases} \tag{6}$$

Таким образом, долю энергий отрезка сигнала в любом частотном интервале можно вычислить на основе представления (5), не вычисляя при этом соответствующую трансформанту Фурье.

В работе [7] также показано, что с целью упрощения вычислений можно воспользоваться тем свойством матрицы **A**, что для нее существует N собственных векторов $\vec{q}_k$, которые соответствуют собственным числам λ_k [8].

Вычисления показывают, что при выполнении неравенства $M = 2[N(\upsilon_2 - \upsilon_1)/2\pi] \geq 4$ собственные числа обладают следующими свойствами

$$\lambda_1 \approx \lambda_2 \approx \lambda_3 \approx \lambda_4 \approx ... \approx \lambda_M \approx 1;$$
$$\lambda_{J+k} \approx 0, \quad k = 1,\ 2,\ ...,$$

где $J = M + 2$.

Представление (5) нетрудно преобразовать к виду

$$P_V(\vec{x}) \cong \sum_{k=1}^{J} (\alpha_k)^2 , \tag{7}$$

где

$$\vec{\alpha} = \sqrt{L_1} Q_1^T \vec{x} = (\alpha_1, ..., \alpha_N)^T , \tag{8}$$

$Q_1 = (\vec{q}_1, ..., \vec{q}_J)$ – подматрица собственных векторов; $L_1 = diag(\lambda_1, ..., \lambda_J)$ – подматрица собственных чисел матрицы **A**.

Если частотный диапазон разбить на равное количество непересекающихся интервалов $R = \dfrac{\pi}{\upsilon_2 - \upsilon_1}$, то можно составить матрицу

$$AA = \begin{pmatrix} \sqrt{L_1^1}\,(Q_1^1)^T \\ \sqrt{L_1^2}\,(Q_1^2)^T \\ \cdots \\ \sqrt{L_1^R}\,(Q_1^R)^T \end{pmatrix}. \tag{9}$$

Тогда для вычисления полного набора долей энергии отрезка сигнала могут служить соотношения

$$\vec{\alpha\alpha} = AA\vec{x} = (\vec{\alpha}_1, ..., \vec{\alpha}_R)^T ; \tag{10}$$

$$P_{Vr}(\vec{x}) \cong \sum_{k=1}^{J} (\alpha_{kr})^2 . \tag{11}$$

При этом точность вычисления доли энергий отрезка сигнала практически сохраняется на уровне представления (5).

Очевидно, что соотношения (5) и (11) представляют собой новый инструмент, позволяющий вычислять доли энергий отрезков звуковых сигналов в заданном частотном интервале.

2. Сегментация на участке пауза/звук

Сегментация речевого сигнала на участке пауза/звук осуществляется с помощью решающей функции для проверки гипотезы о том, что анализируемый отрезок сигнала соответствует паузе между звуковыми данными (нулевая гипотеза) [4]:

$$S = \max\left(\frac{(\alpha_{1r})^2}{(\alpha_{1r}^{\Pi})} + \frac{(\alpha_{2r})^2}{(\alpha_{2r}^{\Pi})} \right). \qquad (12)$$

Здесь элементы, стоящие в числителе представляют собой значения энергий, вычисленные для каждого частотного интервала анализируемого отрезка сигнала (7) для двух собственных векторов, соответствующих максимальным собственным числам.

Элементы, стоящие в знаменателе, представляют собой математическое ожидание энергии для каждого частотного интервала для сигнала, соответствующего заранее выбранной "паузе-эталону".

$$\alpha_{kr}^{\Pi} = \frac{1}{N_{\text{отр}}} \sum_{i=1}^{N_{\text{отр}}} (\alpha_{kr,i}^{\Pi})^2 \,, \; k = 1, 2; \quad r = 1, \ldots, R, \qquad (13)$$

где $N_{\text{отр}}$ – количество отрезков "паузы-эталона".

Использование максимального значения увеличивает вероятность правильного обнаружения границы пауза/звук.

Если выполняется неравенство

$$S > h \,, \qquad (14)$$

то нулевая гипотеза отвергается, а в противном случае принимается решение о наличии паузы и отрезок кодируется на основе фиксации его начала и, в необходимых случаях, длительности.

Символ h в правой части неравенства означает порог, обеспечивающий заданный уровень вероятности ложной тревоги. Значение порога может быть адаптивно вычислено на этапе обработки сигнала в "паузе-эталоне".

3. Вычислительные эксперименты

В ходе вычислительных экспериментов было обработано большое количество файлов, содержащих речевые данные (более 60 файлов).

В табл. 1 представлены оценки вероятности правильного и ложного обнаружения пауз, а также коэффициент сжатия сигнала за счет кодирования пауз на участке сигнала в 100000 отсчетов при заданных N и R.

Вероятность правильного обнаружения определялась как

$$P_{\text{ПО}} = \frac{M_{\text{по}}}{M_{\text{с}}} \,,$$

где $M_{\text{с}}$ – длина сигнала, соответствующего паузе; $M_{\text{по}}$ – число значений решающей функции, не превышающих порог.

Вероятность ложного обнаружения определялась как

$$P_{\text{ло}} = \frac{M_{\text{ло}}}{M_{\text{с}}} \,,$$

где $M_{\text{с}}$ – длина сигнала, соответствующего звуку; $M_{\text{ло}}$ – число значений решающей функции не превышающих порог.

Таблица 1

Оценка вероятности правильного и ложного обнаружения пауз, коэффициент сжатия ($K_{сжат}$).

№	N	R	$P_{ПО}$	$P_{ЛО}$	$K_{сжат}$
1	60	2	0,98261	0,0012	1,74
2	60	6	0,98801	0,0006	1,74
3	60	10	0,98441	0,0006	1,73
4	60	15	0,99101	0,0007	1,75
5	60	30	0,98381	0,0007	1,73
6	200	2	0,99001	0,0000	1,75
7	200	20	0,99601	0,0000	1,75
8	200	25	0,99201	0,0022	1,75
9	200	50	0,99801	0,0000	1,75
10	200	100	0,99801	0,0000	1,75

Для иллюстрации полученных результатов ниже приведены рис. 1 и 2, на которых изображены границы пауза/речь и речь/пауза.

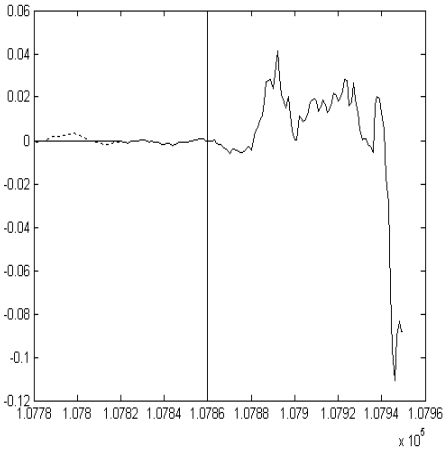

Рис. 1. Граница пауза/звук, определенная при использовании параметров $N = 60$, $R = 10$

При использовании значений параметров $N = 60$, $R = 10$, граница паузы определяется точно, но некоторые короткие участки паузы, чья структура

отличается от структуры сигнала на участке "пауза-эталон" (например, отсчеты с 214600 по 214800 на рис. 2) определяются как речь, что создает "треск" при воспроизведении. Это свидетельствует о чувствительности метода. Так как подобные участки имеют, как правило, малую длительность, то этот эффект можно устранить, например, установив ограничения на длительность участков, соответствующих звуку.

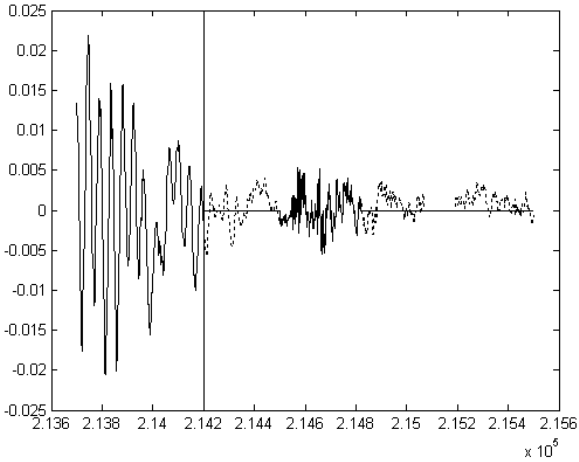

Рис. 2. Граница звук/пауза, определенная при использовании параметров $N = 60$, $R = 10$

В табл. 2 приведена оценка вероятности правильного обнаружения пауз на отрезках сигнала, соответствующих слитной речи. Слова взяты со стечением согласных и содержащие глухие согласные звуки "с", "ф", "ч", которые являются невокализованными звуками, распределение энергий которых подобно распределению энергий пауз.

Таблица 2

Оценка вероятности правильного обнаружения пауз на отрезках сигнала, соответствующих слитной речи

Вероятность правильного обнаружения пауз $P_{\text{ПО}}$ **,%**			
Слово "аспект"	Слово "фактически"	Слово "свойству"	Слово "значит"
98,7	97,3	100	98,5

Вероятности правильного обнаружения в словах "аспект", "фактически" и "значит" не достигают 100 процентов. Это объясняется тем, что в состав этого

слова входят звуки, которые принадлежат к невокализованным звукам малой длительности, вероятность пропуска которых наиболее велика, из-за их малого (по сравнению с вокализованными звуками) уровня, и в данном случае отрезки определенные как паузы приходятся на окончание звука "к" и начало звуков "т" и "ч". Срезание начала звуков в этих случаях особенно нежелательно, так как это может снизить разборчивость речи. Тем не менее, экспертная оценка при воспроизведении сигнала с удаленными паузами показала, что звуки "к", "т" и "ч" в приведенных словах четко различимы.

На рис. 3 представлен отрезок речевого сигнала, включающий как паузы, так и звуки, с длительностью 2,3 секунды и частотой дискретизации 7350 Гц. Сплошной линией показано значение решающей функции. Из рисунка видно, что значения решающей функции значительно превышает пороговое значение на отрезках, соответствующих звукам.

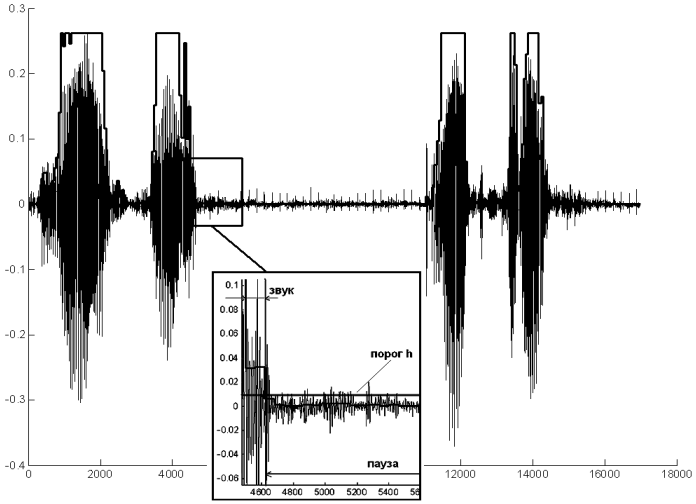

Рис. 3. Отрезок сигнала, соответствующий словосочетанию "свойству спектров", и его решающая функция

Выводы. Предлагаемый алгоритм сжатия речевых данных за счет обнаружения и кодирования пауз на основе сравнения распределений энергии шума и смеси сигнал+шум в заданных частотных интервалах обладает высокой работоспособностью. При всех использованных сочетаниях N и R вероятность правильного обнаружения пауз не менее чем 0,98, а ложного обнаружения пауз не превосходит 0,005. Полученные при этом коэффициенты сжатия имеют значения более 1,7 раза. По результатам вычислительных экспериментов рекомендуется использовать длины анализируемых отрезков $N = 60$ при количестве частотных интервалов $R = 10$, т.к. при этом адекватно

учитываются узость частотных интервалов, где сосредоточена энергия речевых сигналов, и объем вычислительных работ.

Список литературы: 1. *Орищенко В.И.* Сжатие данных в системах сбора и передачи информации / *В.И. Орищенко, В.Г. Санников. В.А. Свириденко*. Под ред. В.А. Свириденко. – М.: Радио и связь, 1985. – 184 с. **2.** *Шелухин О.И., Лукьянцев Н.Ф.* Цифровая обработка и передача речи. – М.: Радио и связь, 2000. – 456 с. **3.** *Жиляков Е.Г., Белов С.П., Прохоренко Е.И.* Новый метод сжатия речевых данных / Труды учебных заведения связи. – СПб. – 2006. – № 175. – С. 152–161. **4.** *Савченко В.В.* Автоматическая обработка речи по критерию минимума информационного рассогласования на основе метода обеляющего фильтра // Радиотехника и электроника. – 2005. – Том 50. – № 3. – С. 309–315. **5**. *Фант Г.* Акустическая теория речеобразования. – М.: Наука, 1964. – 283 с. **6.** Физиология речи. Восприятие речи человеком / *Л.А. Чистович* и др. – М.: Наука, 1976. – 386 с. **7.** *Жиляков Е.Г., Белов С.П., Прохоренко Е.И.* Вариационные методы частотного анализа звуковых сигналов // Труды учебных заведений связи / СПбГУТ. – 2006. – № 174. – С. 163–170. **8.** *Гонтмахер Ф.Р.* Теория матриц. – М.: Физматлит, 2004. – 560 с.

Поступила в редакцію 03. 09. 2007