УДК 681.327.12:534.78

О СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ НА ОДНОРОДНЫЕ ОТРЕЗКИ
ON SEGMENTATION OF SPEECH SIGNALS ON HOMOGENEOUS PIECES

Е.Г. Жиляков, С.П. Белов, А.С. Белов, А.А. Фирсова
E.G. Zhilyakov, S.P. Belov, A.S. Belov, A.A. Firsova

Белгородский государственный национальный исследовательский университет, Россия, 308015, Белгород, ул. Победы, 85
Belgorod State National Research University, 85 Pobeda St, Belgorod, 308015, Russia

e-mail: zhilyakov@bsu.edu.ru,belov@bsu.edu.ru,belov_as@bsu.edu.ru, firsova_a@bsu.edu.ru.

Аннотация. В настоящее время наибольших успехов в решении задачи автоматического распознавания устной речи удается достичь при использовании так называемых скрытых марковских моделей (СММ) речеобразования. Одним из этапов таких процедур является сегментация речевых сигналов (РС) на участки однородности. Вместе с тем, следует отметить, что для использования СММ необходимо иметь достаточно обширные и достоверные сведения о вероятностях переходов из состояния в состояние (звук-звук), а также адекватные модели генерации соответствующих отрезков РС, например, параметров моделей линейного предсказания (ЛП). Поэтому, несмотря на определенное изящество теоретических обоснований и известные примеры общедоступных технологий распознавания устной речи на основе СММ, нельзя считать, что проблема полностью решена. В статье предложен метод сегментации речевых сигналов на отрезки, порождаемые звуками речи, в основе которого лежат частотные представления. Разработаны модели анализа и сопоставления энергетических характеристик отрезков речевых сигналов при принятии решений о значимости их различий.

Resume. At the present time, the greatest success in solving the problem of automatic speech recognition can be achieved by using the so-called Hidden Markov Models (HMM) speech production. One step of such procedures is the segmentation of speech signals (RS) to homogeneity portions. However, it should be noted that the use of CMM must have quite extensive and reliable information on the probability of transition from state to state (audio sound), as well as an adequate model of the generation of the corresponding segments of MS, such as model parameters of linear prediction (LP). Therefore, despite a certain elegance of theoretical studies and known examples of public speech recognition technology based on the CMM, we cannot assume that the problem is completely solved. This paper proposes a method for segmentation of speech signals into segments generated by the sounds of speech, which is based on the frequency of submission. The models of analysis and comparison of energy characteristics of segments of speech signals developed when deciding on the significance of their differences.

Ключевые слова: сегментация речевых сигналов, однородные отрезки речевых сигналов, энергетические характеристики речевых сигналов, анализ и сопоставление, решающее правило.
Keywords: segmentation of speech signals, homogeneous pieces of speech signals, the energy characteristics of speech signals, analysis and collation, decision rule.

Введение

Речевым сигналом (РС) $x(t)$, где (t) - время, будем называть колебания электрического тока (или напряжения) на выходе микрофона, возбуждаемые акустическими воздействиями, которые возникают в процессе информационного обмена на основе устной речи.

Основными элементами устной речи являются ее звуки, комбинации которых (слова) образуют коды различных предметов. Еще одним важным элементом устной речи служат паузы между словами или отдельными звуками.

Возбуждаемые в процессе устной речи акустические колебания дешифрируются в слуховой системе человека, включающей в себя гидромеханические элементы, преобразующие их в электрические импульсы в нейронах мозга, сочетания которых распознаются как слова.

Очевидно, что для обеспечения достоверности в дешифрировании элементов устной речи соответствующие им акустические колебания с одной стороны должны быть в достаточной мере отличаться, а с другой - в пределах их звучания не должны изменяться значимо.

Таким образом, с элементами устной речи можно связать понятие однородности – незначимая в смысле определенной меры изменчивость характеристик возбуждаемых в течение их длительности колебаний акустической среды (переносчика информации).

Ясно, что однородные участки колебаний акустической среды будут порождать однородные в том же смысле отрезки РС. Разбиение РС на однородные отрезки естественно называть сегментацией. Наибольший интерес представляет сегментация РС в автоматическом режиме на основе вычислительных процедур, моделирующих процессы речеобразования и речевосприятия. Ясно, что достижение успеха в автоматической сегментации может служить основой многих задач

обработки речи: автоматического распознавания устной речи, идентификации дикторов, психофизических исследований состояния человека, сжатия речевых данных при хранении и передаче и т.д.

Проблема автоматического анализа и синтеза устной речи на основе обработки РС исследуется достаточно интенсивно [Шелухин, Лукьянцев, 2000]. В частности, задача сегментации РС на однородные отрезки рассматривалась в ряде работ [Сорокин, Цыплихин, 2004; Сорокин, Цыплихин, 2006; Жиляков и др.,2011], где предложены некоторые меры однородности и на основе вычислительных экспериментов исследуется их эффективность. На наш взгляд результаты этих исследований иллюстрируют неадекватность предлагаемых моделей с точки зрения отражения процессов речевого информационного обмена.

По-видимому, в настоящее время наибольших успехов в решении задачи автоматического распознавания устной речи удается достичь при использовании так называемых скрытых марковских моделей (СММ) речеобразования [Рабинер, 1989; Аграновский, Леднов, 2004]. Одним из этапов таких процедур является сегментация РС на участки однородности. Вместе с тем, следует отметить, что для использования СММ необходимо иметь достаточно обширные и достоверные сведения о вероятностях переходов из состояния в состояние (звук-звук), а также адекватные модели генерации соответствующих отрезков РС, например параметров моделей линейного предсказания (ЛП) [Рабинер, 1989]. Поэтому, несмотря на определенное изящество теоретических обоснований и известные примеры общедоступных технологий распознавания устной речи на основе СММ, нельзя считать, что проблема полностью решена.

Целью данной работы является разработка метода сегментации РС на основе моделей анализа энергетических характеристик их отрезков.

Модель анализа энергетических характеристик отрезков РС

Пусть $x(t)$, где $t \in [0.T]$ – отрезок РС, трансформанта Фурье которого

$$X(\omega) = \int_0^T x(t) e^{-j\omega t} d\omega \tag{1}$$

определяет его частотный спектр, так что имеет место равенство Парсеваля:

$$\|x\|^2 = \int_0^T x^2(t) dt = \int_{-\infty}^{\infty} |X(\omega)|^2 d\omega / 2\pi \tag{2}$$

Ясно, что соотношение (2) можно переписать в виде:

$$\|x\|^2 = \sum_{r=0}^{\infty} P_r(x) \ , \tag{3}$$

где $P_r(x)$ - части энергии РС

$$P_r(x) = \int_{\omega \in \Omega r} |X(\omega)|^2 d\omega / 2\pi \ , \tag{4}$$

попадающие в частотные интервалы вида:

$$\Omega_r = [-\Omega_{2r} - \Omega_{1r}) \cup [\Omega_{1r}, \Omega_{2r}), \tag{5}$$

$$\Omega_{1r} = \Omega_{2,r-1}, \Omega_{10} = 0 \tag{6}$$

Представляется адекватным предположение о том, что реакция гидромеханических элементов слуховой системы человека определяется энергией, которая попадает в частотные полосы, соответствующие их упругим свойствам. В качестве основных подтверждений этому можно привести, по крайней мере, два экспериментально установленных факта.

а) Наличие критических частотных полос человеческого слуха, когда два гармонических акустических воздействия на слух не различаются, если их частоты находятся в границах полосы.

б) Основное свойство соответствующих звукам речи отрезков РС заключается в высокой частотной концентрации их энергии, так что найдется такое конечное множество частотных интервалов R_γ, что будут иметь место неравенства:

$$\sum_{r\in R_\gamma} P_r(x) \geq \gamma \|x\|^2 , \tag{7}$$

$$1-\gamma \ll 1' \tag{8}$$

$$\sum_{r\in R_\gamma} |\Omega_{2r} - \Omega_{1r}| \leq \Omega_в , \tag{9}$$

здесь $\Omega_в$ – граничная круговая частота

$$\Omega_в = 2\pi\nu_в \tag{10}$$

$$\upsilon_в \leq 6*10^3 + 10/T \ Гц \tag{11}$$

Следующее предположение модели заключается в том, что при дешифрировании акустических воздействий выделяются частотные интервалы, в которые попадают части энергии, превышающий некоторые пороги:

$$P_r(x) > h_r(x). \tag{12}$$

Представляется естественными такие частотные интервалы называть информационными.

При этом пороги определяются адаптивно, в зависимости от общего фона, в качестве которого естественно использовать спектральную плотность анализируемого отрезка

$$S(x) = \|x\|^2 / (2\Omega_в) \tag{13}$$

Тогда в качестве порогов естественно использовать следующие значения:

$$h_r(x) = 2S(x) * (\Omega_{2r} - \Omega_{1r})_r = \|x\|^2 * (\Omega_{2r} - \Omega_{1r}) / \Omega_в \tag{14}$$

Пусть теперь $R_1(*)$ – множество частотных интервалов, для которых выполняются неравенства вида (12), где правые части определяются соотношениями (14). Положим

$$G_1(x) = \sum_{r\in R_1(x)} P_r(x) , \tag{15}$$

$$D_1(x) = G_1(x) / \|x\|^2 . \tag{16}$$

Последнее соотношение определяет долю энергии отрезка сигнала, попадающую в совокупность выделяемых частотных интервалов. В соответствии с (12) и (14) имеют место неравенства

$$G_1(x) > S_1(x) = \sum_{r\in R_1(x)} h_r(x) \tag{17}$$

$$D_1(x) > \sum_{r\in R_1(x)} (\Omega_{2r} - \Omega_{1r}) / \Omega_в = \Delta\Omega_1(x) / \Omega_в \tag{18}$$

т.е. доля энергии, попадающая в отбираемые частотные интервалы, больше доли соответствующей суммарной частотной полосы:

$$W_1(x) = \Delta\Omega_1(x)/\Omega_в . \quad (19),$$

Отметим, что для отрезков с почти равномерным распределением энергии в пределах полосы $[0, \Omega_в]$ будет иметь место приблизительное равенство

$$D_S(x) \approx W_1(x) \quad (20)$$

Такая ситуация возникает, например, в паузах речи, когда РС представляет собой отрезок шумов микрофона.

В свою очередь в случае вокализованных звуков речи будет реализовываться максимальная частотная концентрация энергии.

Модель сопоставления энергетических характеристик смежных отрезков РС

Разобьем условно отрезок $x(t)$ на два смежных

$$x_1(t) = x(t), \ \ 0 < t \le T/2 , \quad (21)$$

$$x_2'(t) = x(t + T/2), \ \ 0 < t \le T/2 . \quad (22)$$

Спектры этих частей определяются интегралами вида (1) с соответствующими верхними пределами.

Пусть далее $P_r(x)$, $P_r(x_1)$, $P_r(x_2)$ - части энергии общего отрезка и его частей, попадающие в одни и те же частотные интервалы, а множество $\mathbb{R}_1$(x) определяется на основе неравенства вида (12), с учетом (13) и (14).

Представляется адекватным предположение (гипотеза) о том, что изменение энергетических характеристик с течением времени приводит к изменению амплитуды колебаний воспринимающих их элементами слуховой системы, которые и фиксируются интеллектуальной ее частью. При этом принимаются во внимание некоторые средние характеристики колебаний, определяемые спектральной плотностью объединенного отрезка.

Так как энергия гармонического колебания $\alpha\cos(\omega_0 t)$ пропорциональна α^2, то по аналогии с этим полагаем, что амплитуды колебаний гидромеханических элементов для одних и тех же частотных интервалов пропорциональны квадратным корням из соответствующих энергий, то есть $P_r^{1/2}(x_1)$ и $P_r^{1/2}(x_2)$ соответственно.

В качестве меры относительных различий в интенсивностях реакций элементов слуховой системы для объединенного отрезка предлагается использовать функционал:

$$B_1(x_1, x_2) = C_1(x_1, x_2)/\|x\|^2 \quad (23)$$

где,

$$C_1(x_1, x_2) = \sum_{r \in R_1(x)} (P_r^{1/2}(x_1) - P_r^{1/2}(x_2))^2 \quad (24)$$

или, после очевидных преобразований:

$$C_1(x_1, x_2) = \sum_{r \in R_1(x)} (P_r(x_1) - P_r(x_2)) - 2\sum_{r \in R_1(x)} (P_r(x_1) \cdot P_r(x_2))^{1/2} \quad (25)$$

Непосредственно из определения (24) следует неравенство:

$$0 \le B_1(x_1, x_2) \quad (26)$$

Ясно, что равенство нулю здесь достигается тогда и только тогда, когда выполняются все равенства вида:

$$P_r(x_1) = P_r(x_2), r \in R_1(x)^{\cdot} \quad (27)$$

Эти равенства будем называть условиями (признаком) полной однородности отрезка $x(t)$. Такая ситуация вполне возможна для периодического с периодом T/2 сигнала, когда имеет место:

$$x_2(t) = x_1(t) \quad (28)$$

Реальные РС таким свойством не обладают. Поэтому в общем случае необходимо ввести меру значимости отклонения от нуля значений функционала (23), превышение которой принимается за признак нарушения однородности.

В качестве такой меры предлагается использовать:

$$V(x) = S_1(x)/G_1(x) \quad (29)$$

Напомним, что, согласно определениям (15) и (17), знаменатель здесь равен доле энергии объединенного отрезка, попадающей в совокупность информационных частотных интервалов, удовлетворяющих неравенству вида (12), тогда как числитель равен сумме вычисляемых на основе представления (14) порогов. Последнее соответствует среднему значению энергии отрезка, которую можно соотнести с набором информационных частотных интервалов.

Имея в виду определения (14), (15) и (19) границу (29) можно преобразовать к виду:

$$V(x) = W_1(x)/D_1(x), \quad (30)$$

т.е. отношение доли частотной полосы, которую занимают информационные частотные интервалы к доле, попадающей в них энергии объединенного отрезка.

Таким образом, можно сформулировать следующее решающее правило.

Анализируемый отрезок $x(t), t \in [0,T]$ признается однородным при выполнении неравенства:

$$B(x_1, x_2) < V(x) \quad (31)$$

и неоднородным в противном случае. С учетом введенных ранее определений неравенству (31) нетрудно придать иной вид:

$$C_1(x_1, x_2)/S_1(x_1, x_2) < \left\|x^2\right\|/G_1(x). \quad (32)$$

Или

$$C_1(x_1, x_2)/\Delta\Omega(x) < \left\|x^2\right\|/G(x)*S(x). \quad (33)$$

Левую часть этого неравенства можно называть частотной плотностью суммы (24). В соответствии с (33) допускается, что ее значение может быть больше чем спектральная плотность (13) анализируемого отрезка, причем это превышение тем меньше чем больше $D_1(x)$.

Исследования частично финансировались в рамках гранта РФФИ №15-07-01463 и №15-07-01570

Список литературы
References

1. Шелухин, О.И. 2000. Цифровая обработка и передача речи. М., Радио и связь, 456.
Sheluhin, O.I. 2000. Cifrovaja obrabotka i peredacha rechi. M., Radio i svjaz', 456.

2. Сорокин, В.Н. 2004. Сегментация и распознавание гласных. Журнал Информационные процессы, Т.4, №2: 202-220.

Sorokin, V.N. 2004. Segmentacija i raspoznavanie glasnyh. Zhurnal Informacionnye processy, T.4, №2: 202-220.

3. Сорокин, В.Н. 2006.Сегментация речи на кардинальные элементы. Журнал Информационные процессы, Т.6, №3: 177-207.

Sorokin, V.N. 2006.Segmentacija rechi na kardinal'nye jelementy. Zhurnal Informa-cionnye processy, T.6, №3: 177-207.

4. Жиляков, Е.Г. 2011. Сегментация речевых сигналов на основе анализа распределения энергии по частотным интервалам. Научные ведомости Белгородского государственного университета, № 7 (102) выпуск 18/1, серия Информатика: 187-196.

Zhilyakov, E.G. 2011. Segmentacija rechevyh signalov na osnove analiza raspredelenija jenergii po chastotnym intervalam. Nauchnye vedomosti Belgorodskogo gosudarstvennogo univer-siteta, № 7 (102) vypusk 18/1, serija Informatika: 187-196.

5. Рабинер Л.Р. 1989. Скрытые марковские модели и их применение в избранных приложениях при распознавании речи. ТИИЭР т. 77, №2: 86-120.

Rabiner L.R. 1989. Skrytye markovskie modeli i ih primenenie v izbrannyh prilo-zhenijah pri raspoznavanii rechi. TIIJeR t. 77, №2: 86-120.

6. Аграновский, А.В. 2004. Теоретические аспекты алгоритмов обработки и классификации речевых сигналов. М., Радио и связь, 164.

Agranovskij, A.V. 2004. Teoreticheskie aspekty algoritmov obrabotki i klassifikacii rechevyh signalov. M., Radio i svjaz', 164.