МИКРОСЕГМЕНТЫ КАК ОСНОВНЫЕ ЭЛЕМЕНТЫ ПЕРВИЧНОЙ СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ

В.Г. РУДАКОВ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР
АН СССР, МОСКВА

В.Н. ТРУНИН-ДОНСКОЙ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР
АН СССР, МОСКВА

В докладе рассматривается возможность использования для целей первичной сегментации речевых сигналов микросегментов, определяемых в виде совокупности локальных длительностей. И микросегменты и локальные длительности определяются непосредственно из анализа формы речевой волны во временной области. Использование микросегментов позволяет примерно в 2 раза сократить исходную длительность сигнала для последующего анализа на фонемном уровне, а также указать некоторые параметры фонем.

Известно /I,2/, что вся информация о речевом сигнале содержится в его временной функции $P(t)$, отражающей зависимость звукового давления P на некотором расстоянии от говорящего. Успешному решению ряда проблем анализа речевых сигналов способствует правильное проведение процесса их сегментации/3/. С точки зрения достижения максимальной информативности результатов анализа сегментация должна осуществляться адаптивным способом к последовательным во времени звуковым явлениям /4/. Для первичной сегментации речевой сигнал представляют в виде последовательности вокализованных и невокализованных сегментов/I/. В /3/ в качестве основных элементов первичной сегментации речевого сигнала предложены микрофонемы - участки сигнала на протяжении периода основного тона. Преимуществом такого способа является то, что микрофонемы не связаны со строго постоянным интервалом в I0...20 мс, а также большая уместность микрофонемных и фонемных характеристик, относящихся к одному и тому же диктору и классу звуков речи. Недостатком описываемого способа сегментации является использование для анализа сигнала быстрого преобразования Фурье, обусловленного допущением о квазипериодичности и стационарности сигнала на всём протяжении T_0. Такое допущение к речевому сигналу не совсем справедливо /4/. В /4/ речевой сигнал предложено представлять в виде сложной кривой, а для её анализа - метод разложения сложных кривых на компоненты. Этот метод справедлив для анализа как вокализованных, так и невокализованных сегментов, но он, как и спектральный метод, не предполагает непосредственного анализа формы речевой волны.

Для выявления некоторых параметров, характеризующих форму речевой волны рассмотрим на рис. I фрагмент осциллограммы преобразованного в электрический сигнал $U(t)$ изменения звукового давления $P(t)$ на интервале $[t_0, t_N]$.

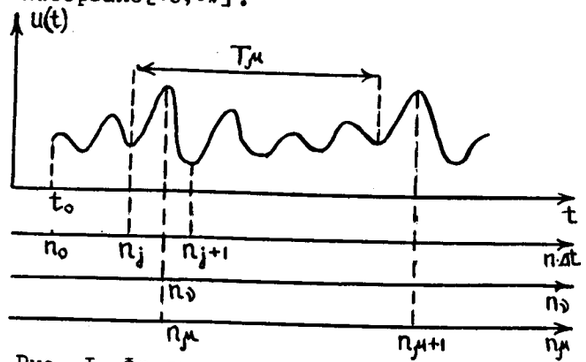


Рис. I. Фрагмент осциллограммы преобразованного в электрический сигнал $U(t)$ речевого сигнала $P(t)$.

Произведем замену непрерывного времени t на дискретное $n\cdot\Delta t$. Принимая $\Delta t = const$, получим зависимость $U(n)$, аргументом которой является номер дискреты $n \in [n_0, n_N]$.

Очевидный колебательный характер функции $U(n)$ можно описать с помощью следующих параметров. На первом уровне описания используются временные интервалы между локальными экстремумами. В номере дискреты n имеет место локальный минимум, если выполнено условие

$$[u(n-1) > u(n)] \wedge [u(n) < u(n+1)],$$

$$u(i) > 0, \; i = n-1, n, n+1. \qquad (1)$$

Для обозначения номеров дискрет, в которых выполняется условие (I) введём индекс j. Смежные локальные минимумы $U(n_j)$ и $U(n_{j+1})$ определяют j-е длительности

$$\tau_j = (n_{j+1} - n_j)\cdot\Delta t = \Delta n_j \cdot \Delta t \qquad (2)$$

Локальные максимумы определяются в пределах изменения τ_j при выполнении условия $[U(n-1) < U(n)] \wedge [U(n) > U(n+1)]$,

$$U(i) > 0, \; i = n-1, n, n+1. \qquad (3)$$

Номера дискрет, удовлетворяющие условию (3) обозначим с помощью индекса ν, тогда локальные максимумы будут иметь обозначение $U^{(j)}(n_\nu)$, где индекс j указывает на их принадлежность к соответствующей τ_j. Длительности τ_ν находятся аналогично выражению (2). Поскольку длительности τ_j и τ_ν определяются соответственно смежными минимумами и максимумами с перекрытием $0{,}5\,\tau_j$ ($0{,}5\,\tau_\nu$), то их наложение может быть использовано как для исключения помех, так и для выявления дополнительных сведений о тонкой структуре сигнала.

Таким образом на первом уровне анализа речевой сигнал представляется последовательностью чисел, характеризующих локальные экстремумы $U(n_j)$, $n_j \in (n_0, n_N)$, $U^{(j)}(n_\nu)$, $n_\nu \in (n_0, n_N)$ и длительности τ_j и τ_ν.

На втором уровне анализа производится выделение значимых экстремумов из локальных. С целью наибольшего учёта динамики функции $P(t)$ необходимо использовать такие однотипные экстремумы, дисперсия которых наибольшая. Проведенный анализ показал, что этому условию удовлетворяют локальные максимумы, поскольку их дисперсия примерно в 4 раза превышает дисперсию минимумов. Значимый максимум определяется из анализа условия

$$[U^{(j-1)}(n_{\nu-1}) < U^{(j)}(n_\nu)] \wedge [U^{(j)}(n_\nu) > U^{(j+1)}(n_{\nu+1})]. \qquad (4)$$

Введём индекс μ для переобозначения таких номеров дискрет n_ν, для которых условие (4) выполняется. Очевидно, что значимый максимум $U^{(j)}(n_\mu)$ всегда совпадает с соответствующим локальным максимумом $U^{(j)}(n_\nu)$.

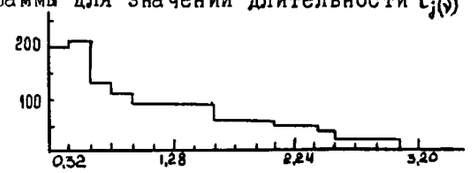
Это позволяет указать временной интервал между смежными значениями максимумов $U^{(j)}(n_\mu)$ и $U^{(j+k)}(n_{\mu+1})$, где k - количество локальных максимумов между номерами дискрет n_μ и $n_{\mu+1}$. Обозначим его через T_μ и определим с помощью выражения

$$T_\mu = (n_{\mu+1} - n_\mu)\cdot\Delta t = \sum_{j=1}^{k+1} \tau_j. \qquad (5)$$

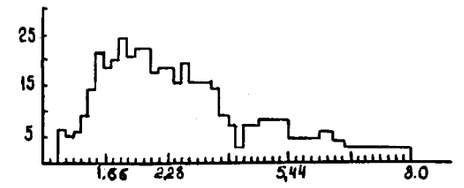
Введём рабочую гипотезу о том, что форма речевой волны в первом приближении может быть охарактеризована параметрами $U(n_j)$, $U^{(j)}(n_\nu)$, τ_j, τ_ν на интервале T_μ, $n_\nu \in [n_j, n_{j+k}]$, $k \in (n_\mu, n_{\mu+1})$.

Для экспериментальной проверки возможности описания формы речевой волны с помощью введённых параметров был использован словарь из 27 слов: ноль, нуль, один, два, три, четыре, пять, шесть, семь, восемь, девять, действие, сложить, вычесть, умножить, величина, точка, цифра, синус, косинус, тангенс, котангенс, слушай, начало, конец, число, целое. Этот словарь по-словно 2-мя мужчинами и женщиной разговорным стилем в помещении машинного зала с уровнем шумов 65 дБ по телефонному каналу с полосой частот 3, I25 кГц, передавался на вход I0-разрядного преобразователя аналог-цифра. В соответствии с указанной полосой частота дискретизации принята 6,25 кГц, что соответствует Δt = I60 мкс.

По результатам обработки 8I слова из указанного словаря на рис. 2 приведены гистограммы для значений длительности $\tau_{j(\nu)}$ и T_μ.



Локальные длительности $\tau_{j(\nu)}$, мс



Длительности T_μ, мс

Рис. 2. Гистограммы длительности $\tau_{j(\nu)}$ и T_μ. Данные получены в результате обработки 8I слова, произнесённых 3-мя дикторами по 27 слов каждый.

Анализ приведенных гистограмм показывает, что локальные длительности $\tau_{j(\nu)}$ занимают диапазон от 320 мкс до 3 мс с максимумом в районе 320 мкс, что соответствует частоте 3,I25 кГц, то есть верхнему значению спектра сигнала. Диапазон значений T_μ находится в пределах от 640 мкс до 8 мс с максимумом на T_μ = 2мс или 500 Гц. Поскольку эти области существенно перекрываются, то они могут быть использованы в ограниченных целях, например, для определения высоты голоса по положению максимума гистограммы длительности T_μ для одного диктора.

Анализ чередований T_μ на протяжении отдельных слов показал, что они обладают определёнными регулярностями. В первом приближении эти регулярности могут быть описаны с помощью семи правил (П1,..П7), которые удовлетворяют следующим выражениям:

$$П1, \; |T_\mu - T_{\mu+1}| \leqslant 160 \text{ мкс} \qquad (6)$$

$$П2, \; |(T_\mu + T_{\mu+1}) - (T_{\mu+2} + T_{\mu+3})| \leqslant 160 \text{ мкс} \qquad (7)$$

$$П3, \; |(T_\mu + T_{\mu+2}) - (T_{\mu+1} + T_{\mu+3})| \leqslant 160 \text{ мкс} \qquad (8)$$

$$П4, \; |T_\mu - (T_{\mu+1} + T_{\mu+2})| \leqslant 160 \text{ мкс} \qquad (9)$$

$$П5, \; |(T_\mu + T_{\mu+1}) - T_{\mu+2}| \leqslant 160 \text{ мкс} \qquad (10)$$

$$П6, \; |(T_\mu + T_{\mu+2}) - T_{\mu+1}| \leqslant 160 \text{ мкс} \qquad (11)$$

$$П7, \; |(T_\mu + T_{\mu+1} + T_{\mu+2}) - (T_{\mu+3} + T_{\mu+4} + T_{\mu+5})| \leqslant \qquad (12)$$

≤ 160 мкс

где 160 мкс = Δt .

Длительности T_μ , удовлетворяющие в своей последовательности соответствующему правилу, объединяются в группы. В речевых сигналах на этом этапе наблюдаются чередования как одинаковых, так и разных групп, которые образуют макрогруппы. Между макрогруппами, а иногда и между группами, встречаются длительности, которые не удовлетворяют приведенным в выражениях (6)...(I2) правилам объединений. Эти длительности не используются для анализа на них распределений $\tau_{j(\nu)}$. Они могут быть учтены лишь при решении вопроса о наличии либо паузы, либо помехи в слове .

Для иллюстрации отмеченных этапов анализа чередования длительностей T_μ на рис. 3 приведена гистограмма объединения T_μ в группы в соответствии с правилами П1,...,П7 в слове "число", диктор мужчина. Из рисунка 3 следует, что каждому правилу объединения T_μ соответствует определённое число групп.

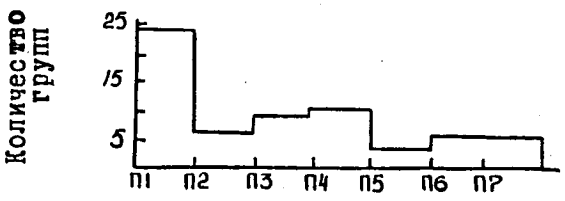


Рис. 3. Гистограмма объединений длительностей в группы с использованием правил П1,...,П7 в слове "число"

Анализ гистограмм слов приведенного выше словаря показал, что для целей сегментации следует выбирать либо такие группы, у которых количество T_μ не менее 7, либо макрогруппы с числом однотипных групп не менее 4 и количеством T_μ в них более 4. Существующую в слове "число" последовательность обозначим в виде следующих групп: П7.I; П7.2; П1.3; П7.4; П7.5; П7.6; П2.7; П2.8; П7.9, где вторая цифра после номера правила объединения обозначает порядок следования групп.

Анализ этих последовательностей показал, что группы П1.3, П2.7 и П2.8 соответствуют вокализованным сегментам, так как у них нет дробления длительностей T_μ. В группе П1.3 они находятся в диапазоне от 1,32 до 2,4 мс, в группе П2.7 - от 2,72 до 3,52 мс, в группе П2.8 - от 2,88 до 3,52 мс. Второй характеристикой этих сегментов является распределение $\tau_{j(\nu)}$. В группе П1.3 длительности $\tau_{j(\nu)}$ в основном находятся в области от 0,32 до 0,96 мс, в П2.7 - от 1,12 мс до 2,56 мс, а в П2.8 - от 1,44 до 2,08 мс. Распределения T_μ и $\tau_{j(\nu)}$ на них в остальных группах характеризуют невокализованные сегменты . Ориентировочно приведенные группы могут быть соотнесены с фонемами: П7.I, П7.3 - "Ч"; П1.3 - "И"; П7.4, П7.5 - "С"; П7.6 - "Л"; П2.7, П2.8 - "О" . Следовательно , по распределениям такого типа можно приблизительно производить сегментацию речевых сигналов на вокализованные и невокализованные участки, а также выносить определённые суждения о фонемных характеристиках выделенных сегментов.

В заключение рассмотрим таблицу; в таблицу сведены соотношения длительностей приведенных групп с длительностью слова.

Из таблицы следует, что общая длительность слова "число" составляет 623,6 мс, а длительность групп - 181 мс, что соответствует 29% от длительности слова. Кроме того, длительности групп: П7.I, П7.2, П7.4, П7.5, П7.6 и П7.9 в среднем совпадают с общепринятым окном анализа в 10...20 мс.

Таблица. Представление длительности слова "число" через длительности групп и интервалов между ними

	Длительность, мс								
Группы	П7.1	П7.2	П1.3	П2.4	П7.5	П7.6	П2.7	П2.8	П7.9
Длительности групп	12,9	8,3	28	15,8	18	16,6	42,8	32,8	13,9
Интервалы между группами	70	43,6	69,1	43,3	0	42,5	0	166	

Всё это позволяет сделать вывод о целесообразности использования в качестве основных элементов первичной сегментации речевые сигналы с длительностями T_μ , которые предлагается называть микросегментами.

Л И Т Е Р А Т У Р А :

I. Фант Г. Акустическая теория речеобразования. Пер. с англ. под ред. В.С. Григорьева. - М.: "Наука", 1964, 283 с.

2. Линдсей П., Норман Д. Переработка информации у человека. Пер. с англ. под ред. А.Р. Лурия. - М.: "Мир", 1974, 550 с.

3. Джерниковский А. Микрофонемы как основные сегменты первичной сегментации речевого сигнала. - Автоматическое обнаружение микрофонем. В Трудах IУ Международной объединенной конференции по искусственному интеллекту. Тбилиси, 1975, том 5, с. 68 - 82.

4. Соломатин В.Ф. Метод разложения сложных кривых на компоненты. Деп. ВИНИТИ, № 4967-81, - М.; 1981, 15 с.