

КОНТРОЛЬНЫЙ ЛИСТОК
СРОКОВ ВОЗВРАТА

КНИГА ДОЛЖНА БЫТЬ
ВОЗВРАЩЕНА НЕ ПОЗДНЕЕ
УКАЗАННОГО ЗДЕСЬ СРОКА

Колич. пред. выдан.

27 ФЕВ 1998

08 АПР 2000

10

22 АПР 1998

3 ТМО Т. 3.600.000 З. 852—87

АКАДЕМИЯ НАУК БЕЛАРУССКОЙ ССР
ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
ИНСТИТУТ ТЕХНИЧЕСКОЙ КИБЕРНЕТИКИ

АНАЛИЗ И СИНТЕЗ РЕЧИ

Сборник научных трудов

Научный редактор доктор технических наук Б.М.Лобанов

Минск 1991

1р.10к.

АКАДЕМИЯ НАУК БЕЛОРУССКОЙ ССР
ОРДENA ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
ИНСТИТУТ ТЕХНИЧЕСКОЙ КИБЕРНЕТИКИ

АНАЛИЗ И СИНТЕЗ РЕЧИ

Минск 1991

М.211198

31.1

А 64

УДК 681.5.01:621.395

М-21198

СОВЕТСКАЯ
ПРАВЯНАЯ
БИБЛИОТЕКА

Представлены работы по анализу и синтезу речи, выполненные в лаборатории автоматического распознавания и синтеза речи Института технической кибернетики АН БССР. В статьях сборника отражены результаты теоретических разработок, проведенных в последние годы, и итоги экспериментальных и конструкторских работ по внедрению новых методов анализа и синтеза практики. В связи с проведенными разработками появляется возможность использования синтезатора речи для озвучивания текстовой информации в персональном компьютере, при создании тренажеров в авиации и на автомобильном транспорте, в области компьютерного обучения языкам, при создании АРМов для инвалидов по зрению.

Сборник будет полезен для специалистов, занимающихся речью, для кибернетиков, программистов, конструкторов РСА, связистов, филологов, эргономистов.

Печатается по решению редакционно-издательского совета Института технической кибернетики АН БССР.

ВВЕДЕНИЕ

Речевой способ передачи деловой информации, по данным специалистов, изучающих взаимодействие людей в производственных условиях, используется на порядок чаще, чем другие способы передачи информации – рукопись, машинопись, телефон. Велико повому стремление специалистов, работающих над созданием новых информационных технологий, систем искусственного интеллекта и автоматизированных систем различного назначения, оснастить реализующие их вычислительные комплексы средствами речевого ввода-вывода информации.

Однако до последнего времени широкое внедрение речевых средств общения человека с ЭВМ сдерживается рядом существенных ограничений, присущих современным системам распознавания и синтеза речи. Для систем распознавания речи это прежде всего невозможность речевого ввода в привычной для человека форме слитной речи, необходимость обучения системы на голос каждого нового пользователя, низкая помехоустойчивость канала речевого ввода информации. Для систем синтеза речи – недостаточная натуральность синтезированной речи, невозможность задания требуемого множества голосов. Весьма актуальными остаются и вопросы технологичности устройства речевого ввода-вывода информации, снабжения их прикладным математическим обеспечением.

В этой связи представляют повышенный интерес работы, направленные на устранение названных выше ограничений в разрабатываемых сегодня системах анализа, распознавания и синтеза речевых сигналов. Этот критерий явился определяющим при отборе работ для настоящего сборника. Насколько успешно нам удалось это сделать – судить читателям.

ISBN 5-7815-0828-7

© Институт технической
кибернетики АН БССР,
1991

АНАЛИЗ И СИНТЕЗ РЕЧИ

ЧАСТЬ I. АНАЛИЗ И РАСПОЗНАВАНИЕ РЕЧЕВОГО СИГНАЛА

УДК 621.391

Н.П. Дегтярев

ПАРАЛЛЕЛЬНО-ПОСЛЕДОВАТЕЛЬНАЯ МОДЕЛЬ АНАЛИЗА, ОБНАРУЖЕНИЯ И ИНТЕРПРЕТАЦИИ СИГНАЛОВ СЛИТНОЙ РЕЧИ

Новые возможности, которые предоставляет речевой способ управления в технических системах уже два десятилетия у нас в стране [1,2] и за рубежом [3,4], стимулируют интенсивные исследования в области разработки систем автоматического распознавания и синтеза речевых сигналов. И тем не менее, если в области разработки систем синтеза речи достигнуты результаты, пригодные для широкого промышленного использования [5,6], то в области автоматического распознавания речи, несмотря на широкий фронт работ и их глубину, удалось лишь продвинуться в понимании сложности решаемой проблемы [7-10], наметить новые направления работы [1,8,II,12] и получить весьма скромные практические результаты [3,4]. К настоящему времени в области распознавания речи сложилось такое положение, когда исследования идут одновременно в нескольких параллельных направлениях, ни одно из которых пока не может продемонстрировать свое преимущество. В условиях сложившегося паритета требуется выработка такого подхода, который позволил бы суммировать положительные свойства известных моделей и стать основой для разработки промышленной системы распознавания речи. В качестве такой основы предлагается параллельно-последовательная модель анализа, обнаружения и интерпретации речевых сигналов, допускающая решение задачи распознавания и понимания слитной речи многих дикторов в условиях воздействия акустических помех.

I. Модуляционная основа инвариантов акустического описания артикуляции речи

Известно [13], что минимальным речеобразующим жестом является слог, минимальным смыслообразующим элементом речи является слово, а минимальным смысловым элементом речевого сообщения является предложение. В общем случае речевое сообщение состоит из последовательности слов и образуется путем артикуляции последовательности составляющих их слогов. В предельном случае речевое сообщение может состоять из одного односложного слова и образовываться минимальным артикуляционным жестом – одним слогом. Таким образом, любое, даже самое минимальное речевое сообщение может быть образовано только путем непрерывных во времени и упорядоченных в соответствии с заданной программой речеобразования движений артикуляторных органов речевого тракта. Известно также, что в процессе речеобразования вследствие инерционности органов артикуляции имеет место явление коартикуляции соседних звуков речи, т.е. взаимовлияние артикуляции соседних звуков, приводящее к взаимозависимости их артикуляторных параметров.

Отмеченные свойства речеобразования указывают на то, что основой образования речевых сообщений являются закономерные и взаимозависимые движения органов артикуляции, задаваемые программой реализации артикуляторного жеста как минимального смыслообразующего элемента речи. Можно предположить поэтому, что именно закономерности изменений функций движения органов артикуляции (кончика языка, поперечного и продольного перемещения тела языка и др.) в контексте реализации слова как артикуляторного жеста или последовательности артикуляторных жестов и являются объемными инвариантными параметрами описания смысловых элементов речи на артикуляторном уровне. Поскольку процессы артикуляции речи на акустическом уровне отображаются в закономерных изменениях во времени структуры и значений формантных параметров речевого сигнала, то именно в закономерностях изменений формантных параметров речевого сигнала и нужно искаать акустические инварианты описания речевых сообщений. При этом нужно помнить, что искомые инварианты имеют смысл только в контексте различных реализаций одного и того же артикулятор-

ного жеста (слова).

Тогда в качестве инвариантной по дикторам функции $P''(t)$ описания смысловых модуляций параметра $P(t)$, адекватной модуляционным свойствам процесса речеобразования, физиологическому закону восприятия раздражений (закон Бебера-Фехнера) и оценке количества информации, может служить функция

$$P''(t) = \ln P(t) - \ln P(t-\tau) - \ln \frac{P(t)}{P(t-\tau)}, \quad (1)$$

где τ – интервал времени, соответствующий разрешающей способности слуха во времени. Нетрудно видеть, что функция (1) не претерпевает существенных изменений при медленном по сравнению с τ изменениях параметра $P(t)$, когда $P(t) \approx P(t-\tau)$.

При условии, что параметр $P(t)$ явным образом отображает движения артикулятора или изменения во времени формантового параметра, функция $P''(t)$ приобретает смысл фонетической, поскольку она инвариантна относительно средних значений $\bar{P}(t)$, которые характеризуют индивидуальные свойства параметра $P(t)$ артикуляции данного диктора. Поэтому определенным продвижением на пути получения инвариантного по дикторам (1) акустического описания речевых сообщений является разработка двухформантной модели акустического описания артикуляции речи [14, 15]. Необходимо отметить, что преобразование спектральных параметров речевого сигнала по (1) [16] не устраивает дикторской вариативности спектра, связанный с индивидуальным для каждого диктора положением формантных максимумов спектра по оси частот.

2. Двухформантная модель акустического описания артикуляции речи

Известно, что каждому из способов образования звуков речи свойственна своя акустическая и эквивалентная ей формантная модель. Поэтому структура модели и набор значащих формантных параметров различны для различных способов образования. Следовательно, решение задачи формантового анализа речевого сигнала возможно только в рамках управляемой модели, структура которой изменяется соответственно с текущим способом образования, гипотезируемым с верхним уровнем распознавания. Это значит, что детализированное формантное описание артикуляции речи по своей

6

природе вторично, так как оно обусловлено определенным фонетическим контекстом. Но тогда должно существовать некоторое безусловное и в этом смысле первичное описание артикуляции, с которого начинается иерархический процесс распознавания. Двухформантная модель получения первичного акустического описания артикуляции речи [14, 15] не требует гипотезирования фонетического контекста и в то же время хорошо отображает связанные с ним формантные свойства мгновенного спектра речи.

Для построения системы параметров первичного акустического описания артикуляции обратим внимание на общие свойства формант речи, устойчиво проявляющиеся на огибающей спектра. Оказывается, что можно выделить четыре вида огибающих спектра, отображающих основные формантные свойства артикуляции способов образования звуков речи (рис. I). Как видно из рис. I, отдельные форманты или формантные группы проявляются в виде концентрации мощности спектральных отсчетов в определенных частотных областях. Каждая из таких концентраций мощности спектральных отсчетов может быть описана интегральными параметрами

$$A = \sum_{j=1}^m \alpha_j / m; \quad (2)$$

$$F = \sum_{j=1}^m F_j \alpha_j / \sum_{j=1}^m \alpha_j; \quad (3)$$

$$B = \sum_{j=1}^m (F_j - F_{j-1}) \alpha_j / m \max \alpha_j, \quad (4)$$

где α_j – отсчеты мгновенного спектра мощности на частотах F_j ; $j = 1, 2 \dots n$; n – число отсчетов спектра по частоте; $y = \arg Z_j$; $Z_j = \text{sign}(a_j - k \max a_j); 0 < k < 1$; m – число отсчетов, превышающих порог $k \max a_j$.

Свойства параметров (2)–(4) существенно зависят от значения коэффициента k . При $k=1$ параметры A и F представляются амплитудой и частотой максимального отсчета спектра, а при $k=0$ соответственны интенсивность и средней частотой, выраженные через моменты нулевого и первого порядка от отсчетов спектра. Параметр B связан с эффективной шириной спектра, а значение коэффициента k определяет степень его чувствительности к модуляциям пирины спектра. Кроме того, коэффициент k влияет на число отсчетов m , группирующихся около максимального отсчета

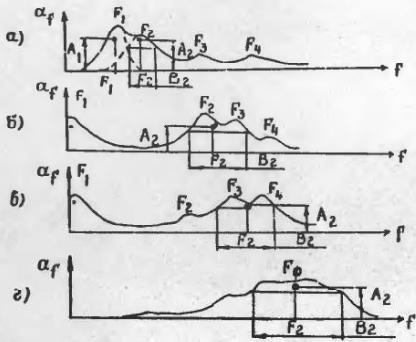


Рис.1. Формантные свойства артикуляции способов образования звуков речи:
а) компактных и аспиративных, б) диффузных,
в) носовых и звонких фрикативных,
г) глухих фрикативных

и определяющих значения и интегральные свойства параметров (2)-(4). Поэтому значение коэффициента k оптимизируется для каждой из двух аппроксимируемых формант групп в зависимости от амплитудных отношений составляющих их формант. Выбор формантных групп и основные принципы обработки их спектрального представления сводятся к следующему.

1. В частотных границах первой форманты звонкие звуки беди образуют формантные группы, состоящие не более чем из двух первых формант. При этом амплитуда первой форманты в таких группах, как правило, является наибольшей. Поэтому выбор значения k_1 в пределах 0,5-0,8 обеспечивает хорошую корреляцию параметров A_1 и F_1 , определяемых по (2) и (3), с амплитудой и частотой первой форманты. Названное выше свойство первой форманты позволяет также обнаруживать первую формантную группу по максимальному отсчету спектра.

8

2. По найденному значению F_1 на втором этапе производится инверсная фильтрация спектра первой форманты, что обеспечивает эффективное разделение первой и второй формант в случаях, когда они составляют одну формантную группу.

3. На третьем этапе полученные спектральные отсчеты описываются параметрами A_2 , F_2 и B_2 , определяемыми соответственно (2) - (4). При выборе значения коэффициента k_2 в пределах 0,3-0,6 названные параметры хорошо отображают амплитудно-частотные отношения второй и более высоких голосовых формант и частотное положение и эффективную ширину фрикативных и аспиративных формант.

На рис.2 приведен пример результата автоматического выделения с помощью модифицированного алгоритма предложенных параметров описания артикуляции. Модификация алгоритма состоит в том, что параметры A_1 и A_2 преобразованы в параметры AT и $A_{\text{Ш}}$, несущие смысл уровней тональных и шумовых сегментов речевого сигнала:

$$AT = \sqrt{A_1 \cdot A_2};$$

$$A_{\text{Ш}} = \begin{cases} A_2 - A_1 & \text{при } A_2 > A_1 \\ 0 & \text{при } A_1 > A_2. \end{cases}$$

Таким образом, на рис.2 представлены в виде функций времени акустические параметры описания артикуляции (в порядке следования снизу вверх) $F_1(t)$, $F_2(t)$, $B_2(t)$, $AT(t)$, $A_{\text{Ш}}(t)$.

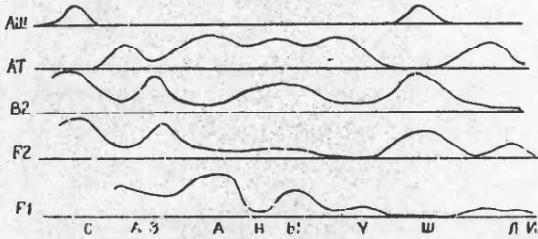


Рис.2. Акустические параметры описания артикуляции речи

9

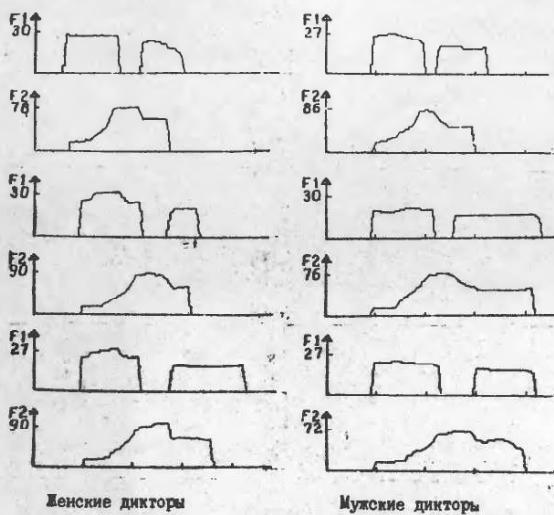


Рис. 3. Закономерности артикуляции слова "ОСЕНЬ" в пространстве вычисляемых по [14] параметров F_1 и F_2 для женских и мужских дикторов

10

Предложенная система акустических параметров первичного описания артикуляции речи составляет хорошую основу для последующего анализа на верхних уровнях модели распознавания ввиду следующего. Во-первых, предлагаемые параметры хорошо отображают формантные свойства спектра речи, во-вторых, они удовлетворяют требованиям линейной модели аппроксимации параметров описания слитной речи [17] и, в-третьих, в рамках названной модели возможно описание топологических инвариантов в смысле преобразования (1) тем самым продвижение в решении проблемы многоодикторности распознавания.

Иллюстрацией последнего может служить пример (рис.3) поведения во времени частот F_1 и F_2 , вычисленных с помощью предложенного алгоритма, для реализаций слова "ОСЕНЬ", произнесенных женскими и мужскими дикторами.

3. Обнаружение и оценка параметров периодичности речевого сигнала

Полная система параметров описания артикуляции речи должна содержать параметры голосового возбуждения. Ниже рассматривается метод обнаружения и оценки параметров периодичности речевого сигнала, основанный на максимизации выходного эффекта приемника различия периодических и шумовых сигналов [18]. Алгоритм вытекает из структуры приемника различия для сигналов при их спектральном представлении [19]

$$P_n = P \left\{ \int_0^{\omega} G(\omega) [S_n(\omega) - S_w(\omega)] d\omega < 1/2 (G_n - G_w) \right\},$$

где P_n – вероятность ошибки обнаружения периодического сигнала; $G(\omega)$ – спектр мощности сигнала на входе приемника; $S_n(\omega) - S_w(\omega)$ – разность эталонных спектров периодического и шумового сигналов; $G_n - G_w$ – разность мощностей периодического и шумового сигналов.

Задача состоит в том, чтобы найти эталонную разность спектров, такую, которая бы максимизировала выходной эффект приемника для периодических и минимизировала его для шумовых сигналов.

Представим описание спектра в виде единичных отсчетов с

11

шагом дискретизации $\Delta\omega$:

$$S(\omega) = \sum_{k=1}^m \delta(\omega - k\Delta\omega); k = \overline{1, m}.$$

Тогда в рамках принятой модели эталонные спектры шумового и периодического сигналов можно представить в виде гребенчатых функций

$$S_w(\omega) = \sum_{k=1}^m \delta(\omega - k\Delta\omega);$$

$$S_p(\omega) = \sum_{k=1}^m \delta(\omega - 2k\Delta\omega),$$

где $\Delta\omega = \pi f_0 T_0$ – частота основного тона периодического сигнала. Выходные эффекты для шумового и периодического сигналов по условиям оптимизации в пределе должны быть равны:

$$\sum_{k=1}^m G_w(\omega) \left[\sum_{l=1}^{m+1} \delta(\omega - 2k\Delta\omega) - \sum_{l=1}^m \delta(\omega - k\Delta\omega) \right] k\Delta\omega = 0; \quad (5)$$

$$\sum_{k=1}^m G_p(\omega) \left[\sum_{l=1}^{m+1} \delta(\omega - 2k\Delta\omega) - \sum_{l=1}^m \delta(\omega - k\Delta\omega) \right] k\Delta\omega = G_p, \quad (6)$$

где $G_w(\omega)$ и $G_p(\omega)$ – спектры мощности шумового и периодического сигналов на выходе приемника различения.

Нетрудно видеть, что условиям (5), – (6) удовлетворяет эталонная разность:

$$S_p(\omega) - S_w(\omega) = \sum_{k=1}^{m+1} \delta(\omega - 2k\Delta\omega).$$

Следовательно, выходной эффект приемника различения может быть выражен

$$G = \sum_{k=1}^{m+1} G(\omega) \left[\sum_{l=1}^{m+1} \delta(\omega - 2k\Delta\omega) - \sum_{l=1}^m \delta(\omega - (2k-1)\Delta\omega) \right] k\Delta\omega.$$

Из последнего выражения следует, что алгоритм различения состоит в суммировании спектральных отсчетов из частот гармоник периодического сигнала и вычитании из этой суммы спектральных отсчетов на частотах, расположенных между гармони-

ками периодического сигнала. Очевидно, что при изменении частоты основного тона необходимо составить ряд указанных сумм для выбранной сетки частот $F_{0i}, i = \overline{1, J}$. Тогда алгоритм оценки частоты основного тона выглядит следующим образом:

$$F_{0i} = \frac{1}{\pi} \arg \min_{\omega} \sum_{k=1}^{m+1} G(\omega) \left\{ \sum_{l=1}^{m+1} \delta(\omega - 2k\Delta\omega_l) - \sum_{l=1}^m \delta(\omega - (2k-1)\Delta\omega_l) \right\} k\Delta\omega_l. \quad (7)$$

Итак, метод оценки частоты основного тона основывается на оценке меры дискретности $G_{0i} = \sum_{k=1}^{m+1} G(\omega)$ спектра сигнала на индикации расстояния F_{0i} между гармониками обнаруженной дискретности. Нетрудно показать, что анализируемый метод эквивалентен автокорреляционному методу обнаружения и оценки параметров периодичности речевого сигнала. Для упрощения реализации метода использовалась огибающая речевого сигнала, что позволило сместить область анализа в диапазон частот основного тона. Результатом оценки частоты основного тона является выбор по (7) значения частоты из установленной дискретной сетки частот.

4. Постановка и решение задачи обнаружения и интерпретации сигналов слитной речи

У нас в стране [20, 21] и за рубежом [22, 23] известно решение задачи распознавания слитной речи, составленной из слов заданного словаря. Постановка и решение задачи в рамках этой концепции состоит в следующем.

Принятая реализация речевого сигнала $A = a_1, a_2, \dots, a_s, \dots, a_j$ аппроксимируется такой последовательностью связанных эталонов $B = B^{n(1)} \oplus B^{n(2)} \oplus \dots \oplus B^{n(x)} \oplus \dots \oplus B^{n(k)}$, которая доставляет минимум расстояния $D(A, B)$ между реализацией A и последовательностью эталонов B . Другими словами, решение сводится к минимизации функционала

$$T = \min_{k, n(x)} D[A, B^{n(1)} \oplus B^{n(2)} \oplus \dots \oplus B^{n(x)} \oplus \dots \oplus B^{n(k)}], k = \overline{1, k},$$

который, в свою очередь, может быть преобразован к следующему структурному виду:

$$T = \min_{k, n(x), l_x} \sum_{i=1}^k D[A(i_{x-1}, i_x), B^{n(x)}], \quad (8)$$

где $\pi(x)$ - исходная последовательность слов (эталонов) в фразе;

k - число слов в фразе; i_x - границы между словами в реализации фразы, для которых выполняется условие

$$i_1 < i_2 < \dots < i_x < \dots < i_{k-1} < j. \quad (9)$$

Оптимизационная задача (8) решается с помощью метода динамического программирования. Нетрудно заметить, что решение задачи распознавания слитной речи согласно (8) и (9) является корректным в тех случаях, когда высказывание, подлежащее распознанию, состоит только из слов заданного словаря. На практике же мы чаще всего сталкиваемся с такими ситуациями, когда распознаваемые фразы содержат не только "свои", но и "чужие" слова, не входящие в заданный словарь, а также помехи различного рода. Иложенная стратегия распознавания не рассчитана на такие случаи, поэтому ответы, получаемые для них путем решения задачи (8), неизбежно будут исказаться, поскольку в минимизируемое интегральное расстояние $D(\cdot)$ будут вносить свой вклад и те сегменты входной реализации (помехи, чужие слова), для которых нет эталонов в заданном алфавите.

Более адекватной реальной стратегией распознаваемых сигналов представляется стратегия обнаружения слов заданного словаря по мере их реализации во входном высказывании [24-26]. При этом становится возможным учесть тот факт, что распознаваемые слова могут следовать друг за другом непрерывно или их могут разделять "чужие" слова или паузы, в том числе и заполнительные помехами различного рода. В таких условиях ответ распознавания естественно искать в виде последовательности $\pi(x)$ только тех эталонов, которые обнаруживают оценки мер сходства $D(\cdot)$ с соответствующими им сегментами реализации $A(i_u^h, i_v^k)$; удовлетворяющие критерии достаточного правдоподобия $r_h(x)$:

$$\pi(x) \in \{\pi(u)\} \exists D(\pi) < r_h(x); x = \overline{i_u N}; \quad (10)$$

Далее из мер $D(\cdot)$ можно составить меры $D(\pi, v)$ сходимости гипотезируемых последовательностей эталонов $\pi(z, v)$, которые обнаруживают достаточное правдоподобие (10):

$$D(\pi, v) = \frac{1}{k} \sum_{i=1}^k D(\pi_i, v). \quad (11)$$

Для каждой последовательности $\pi(z, v)$ выполняется условие чередования границ (начала i_z^h, v и конца i_z^k, v) составляющих ее слов:

$$i_1^h < i_2^k < \dots < i_{z-1}^h < i_z^k < \dots < i_{z-1}^h < i_z^k < v. \quad (12)$$

В (10) - (12) обозначено: k - число слов в з последовательности; $i = \overline{i_z^h, i_z^k}$ - номера отсчетов вектора A параметров описания реализации входного сигнала; $D(\cdot)$ - нормированная K -лине реализации мера сходства, определяемая методом динамического программирования [27] по алгоритмам [24-26]. В процессе вычисления меры сходства находятся границы реализации v -го слова, определяемые как начало i_z^h и конец i_z^k отрезков входного сигнала, "достаточно" близкого к v -му эталону.

Исследуемый алгоритм реализуется в виде трех основных этапов, на каждом из которых производится сокращение исходного множества эталонов по критериям акустического подобия. На первом этапе по текущим оценкам меры близости начальных отсчетов

эталонов формулируется гипотеза о возможном начале i_z^h v -го слова и одновременно из исходного множества $\pi(x)$ эталонов отбираются эталоны, наиболее "похожие" своими начальными отсчетами на текущий отрезок входного сигнала. На втором этапе отбираются эталоны $\pi(v)$, "похожие" на текущий отрезок речевого сигнала всеми своими отсчетами. Одновременно для каждого v -го эталона формируется гипотеза i_z^k об окончании его "похожести" на текущий отрезок речевого сигнала. На третьем этапе из числа эталонов, удовлетворяющих критерию достаточного правдоподобия в каждый момент времени i_z^k , формируются 5 ранжированных по $D(\pi, v)$ последовательностей $\pi(z, v)$ эталонов, границы правдоподобия которых удовлетворяют условию (12). Процесс формирования последовательностей $\pi(z, v)$ заканчивается, как только будет обнаружена межразделенная пауза. Полученные к этому моменту последовательности $\pi(z, v)$ составят набор рабочих гипотез возможных последовательностей слов во фразах или иначе - возможные варианты лексической интерпретации входного высказывания. Полученный набор лексических гипотез может быть подвернут дальнейшему анализу по синтаксическим и семантическим критериям.

В процессе формирования рабочих гипотез имеется возможность сокращения их числа по критерию акустического правдоподобия на каждом шаге этого процесса в i_y^k -е моменты времени. Для этого в каждый i_y^k -й момент времени формируются интегральные меры близости $D_q^n(s, v)$ каждой текущей q последовательности. Причем для каждой из s последовательностей процесс ее формирования продолжается только в том случае, если выполняется условие

$$D^n(s, v) \leq h \min_q D_q^n(s, v); 0 < h < 1; q = \overline{1, q}.$$

При необходимости возможно принятие однозначного решения по критерию наилучшего правдоподобия:

$$n(v) = \arg \min_s D^n(s, v).$$

Итак, стратегия (8) и (9) состоит в аппроксимации всех элементов входного сигнала элементами алфавита распознавания. Это весьма жесткое требование, ибо для его выполнения мы должны уметь синтезировать эталонные сигналы для всего многообразия как сигналов слитной речи, так и помех, что сегодня остается пока еще не решенной проблемой. Напротив, концепция обнаружения заданных элементов (I0), (II) и (I2) требует адекватного описания эталонов слитной речи ограниченного алфавита элементов распознавания, что является более реальной задачей. К тому же, по мере расширения алфавита распознавания, мы, оставаясь в рамках концепции обнаружения заданных элементов, имеем возможность последовательно продвигаться ко все более полному распознаванию входных высказываний, а следовательно, и к более подробной их лексической интерпретации. И, наконец, такая стратегия позволяет эффективно отвергать любые сигналы, "не похожие" на эталонные элементы распознавания, и тем самым повысить помехоустойчивость распознавания. Все это убеждает нас отдать предпочтение последней из рассмотренных нами двух стратегий распознавания слов слитной речи.

.5. Формирование словаря эталонов слитной речи

Для успешного решения проблемы распознавания слитной речи в равной мере важно не только выбрать правильную стратегию распознавания, но и решить задачи, связанные с формированием

словаря эталонов слитной речи. К ним мы относим, прежде всего, выбор базовых элементов словаря и их описание.

В споре о том, какого уровня элементы речи, фонемного или лексического, должны составлять базовый словарь распознавания, мы ищем ответ в третьем решении – базовый словарь распознавания должен состоять из элементов ч фонемного, и лексического уровней, но в определенных отношениях между ними. При параметрическом описании лексических элементов речи принципиальным аргументом против их использования в качестве базовых единиц распознавания считается необходимость обучения системы не только ч словарь распознавания, но и на голос диктора. Однако при описании лексических элементов модуляционными (I) параметрами артикуляции (2) – (4) необходимость обучения на голос диктора упадает, и тем самым снимается принципиальное препятствие на пути использования их в качестве базовых элементов словаря распознавания. Заметим к тому же, что предлагаемое двухуровневое описание лексических элементов речи не противоречит так называемой моторной теории восприятия речевых сообщений, согласно которой наряду с символично-фонетическим описанием слов используется и их целостное описание в виде "артикуляторной оболочки" [28]. При этом целостное описание лексических единиц речи модуляционными параметрами артикуляции может играть роль основного, опорного, а символично-фонетическое – роль вспомогательного, обобщенно-информационного описания. С помощью обобщенного описания, например, в символах устойчиво обнаруживаемых способов образования явлений речи может производиться входжение в языковую, выделение лексических и семантических гипотез и даже принятие однозначных решений в случаях отсутствияльтернативных гипотез.

Важно, что символично-фонетическое описание в роли информационного, вспомогательного не обязано быть подробным и тем самым обходится непреодолимые пока трудности на пути решения проблемы фонетической сегментации речевого сигнала. Но еще более важно то, что все закономерности артикуляции звуков в потоке слитной речи, а также все устойчивые варианты слов и словоформ, рождающиеся действием звуковых законов слитной и разговорной речи [29,30], в принципе могут быть адекватно отражены именем.

но в целостном описании лексических единиц речи модуляционными параметрами артикуляции (1) – (4). С использованием названных параметров описания решается также и задача моделирования коартикуляционных связей на стыках эталонов лексических единиц речи [31]. Таким образом, решение задачи формирования словаря эталонов слитной речи становится возможным путем создания баз данных многодикторных эталонов лексических единиц речи, представленных в виде экономного описания модуляционными параметрами артикуляции и полученных посредством одноразового обучения системы распознавания на словарь требуемых вариантов слов и словоформ.

Заметим, что целостное описание эталонов лексических единиц, будучи представленным в экономном виде методами векторного и сегментного квантования [1,32], по экономности соизмеримо с сегментно-фонетическим. Принципиальная разница между ними состоит в том, что в первом случае используемым для описания речевого сигнала сегментам не приписывается никакого фонетического значения, что и обуславливает реализуемость методов векторного и сегментного квантования речевого сигнала. Возможность экономного представления целостного описания лексических единиц речи, с одной стороны, и многодикторность модуляционных параметров артикуляции, с другой, обеспечивают принципиальную реализуемость использования лексических элементов в составе базового словаря эталонов распознавания слитной речи.

6. Параллельно-последовательная модель понимания речи

Общепринятая модель понимания речи имеет иерархическую структуру, содержащую пять–семь последовательных этапов обработки речевого потока [1,7–10]. Последовательность этапов обработки информации в такой модели обычно отображает и последовательность решаемых задач в проблеме понимания речи. Следуя за эволюцией в понимании проблемы, попытаемся выделить из них приоритетные задачи, направленные на моделирование основных свойств восприятия речи.

Первая принципиально важная задача заключается в разработке методов акустического описания артикуляции. В обобщенной модели понимания речи (рис.4) она относится к уровню моделирования акустического интерпретатора артикуляции речи – АКИНАР.

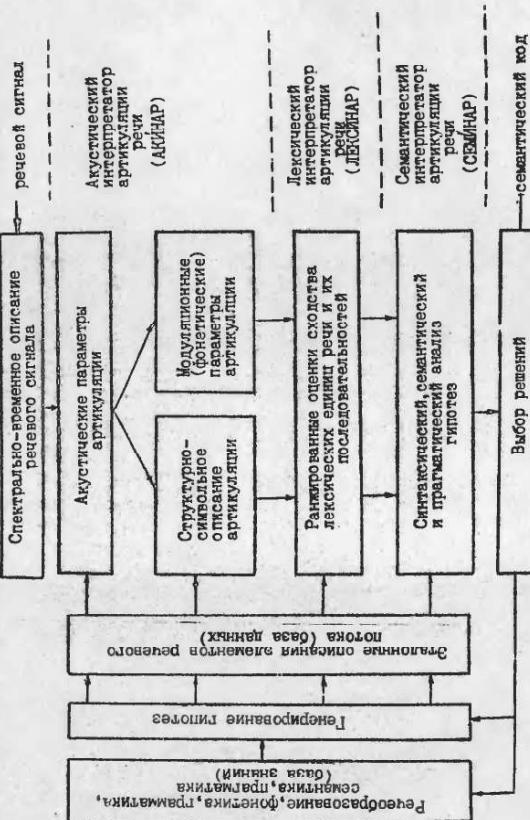


Рис. 4. Текущевневая параллельно-последовательная модель анализа, обнаружения и интерпретации (понимания) сигналов слитной речи

Следующие задачи вытекают из двух важнейших свойств восприятия речевых сообщений. Первое свойство состоит в том, что в отсутствие контекста при восприятии отдельных частей речи (звуков, слов) человек обнаруживает механизм акустического распознавания речевых сигналов. Второе и основное свойство восприятия речи состоит в том, что в процессе речевого общения (диалога) человек на основе акустического распознавания реализует механизм понимания речи путем учета грамматических, семантических и грамматических законов языка. В силу этого надежность понимания речи выше надежности распознавания. Тем не менее, в определенных ситуациях, как мы уже отметили, человек вынужден полагаться только на акустическое распознавание речи. Для нас важно то, что два названных механизма восприятия речи существуют и активно используются как отдельно, так и во взаимосвязи. Отсюда следует, что модель включает в себя еще два уровня обработки информации, соответственно которым формулируются следующие две очевидные задачи. Лексический интерпретатор артикуляции речи – ЛЕСИНАР – решает задачу получения последовательности лексических единиц речи, ранжированных по своим мерам близости к реализации. Ранжированные по мерам близости последовательности лексических единиц речи образуют информационный поток для следующего уровня модели понимания, обозначенного на рис.4 как семантический интерпретатор артикуляции речи – СЕМИНАР.

Для выявления структуры модели обратим внимание на следующее. С одной стороны, на лишен оснований подход к решению проблемы понимания речи через механизм последовательного и подробного анализа речевого потока на всех уровнях иерархической модели [1,7,8]. Но, с другой стороны, параллельно этому механизму обнаруживается и механизм прямого перехода информации на более высокие уровни с пропуском некоторых уровней ее обработки. Например, как уже отмечалось, наряду с символно-фонетическим описанием слов существует так называемое целостное описание и восприятие слова в виде его "артикуляторной оболочки", т.е. имеет место переход от параметрического уровня к лексическому, минуя фонетический. Известно также, что для понимания смысла воспринимаемого сообщения в принципе не требуется распознавание всех элементов сообщения [33]. Это очень важное свойство

речевой коммуникации позволяет человеку успешно воспринимать смысловые сообщения даже в условиях воздействия значительных акустических помех, когда оказывается возможным распознавание лишь отдельных, ключевых элементов сообщения. Следовательно, в таких случаях не работает синтаксический уровень анализа. В связи со сказанным наиболее адекватной естественной модели восприятия представляется так называемая гетерархическая или параллельно-последовательная модель понимания речи [9,34]. Вот почему при постановке и определении приоритетных задач в распознавании речи мы исходим из концепции трехуровневой параллельно-последовательной модели анализа, обнаружения и интерпретации (понимания) сигналов слитной речи (рис.4).

Назанные уровни обработки информации определяют и ключевые задачи в рамках принятой модели. В силу заданной последовательности и тесной взаимосвязи в первую очередь должны решаться задачи артикуляторного и лексического анализа сигналов слитной речи, чemu и удалено основное внимание в настоящей работе. Принципиально важным на наш взгляд являются предложенные в работе подходы к решению актуальных задач артикуляторного и лексического анализа речевых сигналов. Применительно к задаче артикуляторного анализа – это опора на модуляционную и формантную природу инвариантов многодикторного описания артикуляции речи. Применительно к задаче лексического анализа слитной речи – это использование принципа обнаружения эталонов слов слитной речи в текущем речевом сигнале. Последнее важно в том плане, что принцип обнаружения закладывает основу для реализации понимания речевых сообщений через распознавание их отдельных, ключевых лексических элементов. Такой подход согласуется как с механизмами естественного речевосприятия, так и с коммуникативной ролью слов, которые являются "наименееими значимыми элементами, употребляемыми в процессе общения" [29]. Необходимо еще раз подчеркнуть также важность задачи формирования описаний эталонов лексических элементов слитной речи. Можно ожидать, что в ближайшем будущем акцент в проблеме распознавания слитной речи сдвинется в сторону этой задачи.

В практическом плане решение задач артикуляторного и лексического анализа в рамках изложенной концепции позволит повы-

сить понехоустойчивость распознавания и понимания слитной речи многих дикторов и тем самым приблизить способ речевого ввода информации к нормам естественной речи.

ЛИТЕРАТУРА

1. Винцук Т.К. Анализ, распознавание и интерпретация речевых сигналов. - Киев: Наукова думка, 1987. - 264 с.
2. Лобанов Б.М. Исследование и разработка методов автоматического синтеза речи по фонемному тексту. Автореф. дис. ... д.т.н. - Рига, 1984.
3. Современное состояние и тенденции развития устройств речевого ввода-вывода. Приборы, средства автоматизации и системы управления.-Вып. 6. - М.: 1986. - 57 с.
4. Selecting Voice-Recognition Systems. I.C. Magazine, October 27, 1987.-Р. 282-308.
5. Кейтер Д. Компьютеры-синтезаторы речи: Пер. с англ. - М.: Мир, 1985.
6. Кучеров В.Д., Лобанов Б.М. Синтетическая речь в системах массового обслуживания. - М.: Радио и связь, 1983.
7. Бондарко Л.В., Загоруйко Н.Г. и др. Модель восприятия речи человеком. - Новосибирск: Наука, 1968.
8. Методы автоматического распознавания речи. Пер с англ. - М.: Мир, 1983.
9. Галунов В.И. Принципы переработки сложных речевых сообщений. Автоматическое распознавание слуховых образов: Мат. Всесоюз. школы-семинара (APCO-I2). - Киев, 1982. - С. 349-351.
10. Косарев В.А. Естественная форма диалога с ЭВМ. - Л.: Машиностроение, 1989.
11. Левинсон С.Е. Структурные методы автоматического распознавания речи // ТИИР. - Т. 73. - № II. - 1985. - С. 100-120.
12. Слуцкер Г.С., Кринов С.Н. Экспериментальная дикторонезависимая система понимания речи. Речевая информатика. - М.: Наука, 1989. - С. 87-94.
13. Чистович Л.А., Кожевников В.А. и др. Речь. Артикуляция и восприятие. - М. - Л.: Наука, 1965.
14. Дегтярев Н.П. Двухформантная аппроксимация спектров речи//

15. Автоматическое распознавание слуховых образов (APCO-I4).-Ч.1.-Каунас, 1986.-С. 12-13.
16. Дегтярев Н.П. Акустическое описание артикуляции параметрами обобщенных формант спектра речи// Автоматическое распознавание слуховых образов (APCO-I5).-Таллинн, 1989.-С. 145-149.
17. Пирогов А.А. Фонетический код речевого сигнала// Вокодерная телефония.-М.: Связь, 1974.-С. 386-391.
18. Винцук Т.К. О математических моделях речевого сигнала, используемых в распознавании речи// Автоматическое распознавание слуховых образов (APCO-I2): Тез.докл. и сообщений.-Киев, 1982.-С. 34-37.
19. А.С. 581491 (СССР). Способ обнаружения дискретных составляющих спектра речи. Н.П.Дегтярев, Б.М.Лобанов.- Опубл. в ЕИ, 1982.- № 43.
20. Харкевич А.А. Борьба с помехами.-М: Изд-во физ.-мат. литературы, 1963.
21. Винцук Т.К. Распознавание непрерывной речи, составленной из слов заданного словаря // Кибернетика.- 1971.- № 2.- С. 133-143.
22. Винцук Т.К. Система распознавания непрерывной речи "Речь I21". Всесоюз. школа-семинар "Бионика интеллекта".-Харьков, 1987.
23. Sakoe H. Taro-level DP Matching - A Dynamic Programming Based on Pattern Matching Algorithm for Connected Word Recognition, IEEE Журн. со АЗР, 1979. - 27. - № 6.- Р. 588-595.
24. Система распознавания связной речи фирмы ИВО// Зарубежная радиоэлектроника.- 1980.- № 4.- С. 108-120.
25. Дегтярев Н.П. Алгоритм распознавания слов в непрерывном сигнале. Всесоюз. школа-семинар "Бионика интеллекта": Тез. докл.- Харьков, 1987. - С. 27.
26. Дегтярев Н.П., Левков Е.Я. Повышение надежности распознавания слов слитной речи. Proceedings with International Congress of Phonetic Sciences.Vol.3, Tallinn, 1987.-Р.290-293.
27. Дегтярев Н.П., Александровский В.И., Александровская М.И. Исследование алгоритма обнаружения и распознавания слов слитной речи: Тез.докл. и сообщений Всесоюз.школы-семинара (APCO-I5). - Таллинн, 1989.- С. 93-94.

27. Муцкер Г.С. Нелинейный метод анализа речевых сигналов // Труды НИИР. - №2. - 1968. - С. 18-23.
28. Чистович Л.А. Психоакустика и вопросы теории восприятия речи. Опознавание слуховых образов. - Новосибирск: Наука, 1970. - С. 55-141.
29. Гвоздев А.Н. Современный русский литературный язык. Ч.1. Фонетика и морфология. - М.: Просвещение, 1973.
30. Русская разговорная речь / Под ред. Е.А.Земской. - М.: Наука, 1973.
31. Дегтярев И.П. Использование формантно-фонемных связей для формирования эталонов слитной речи: Мат. Всесоюз. школы-семинара "Бионика интеллекта". - Харьков, 1987. - С. 13.
32. Маккоул Д., Рукос С., Гиш Г. Векторное квантование речи // ТИИР. - № II. - 1985. - С. 19-61.
33. Пиотровский Р.Г. Текст, машина, человек. - Л.: Наука, 1975.
34. Мишин Л.Н. Гетерархическая организация распознавания речи. Автоматическое распознавание слуховых образов: Мат. Всесоюз. школы-семинара (APCO-I0). - Тбилиси: Мецниереба, 1978. - С. 101-102.

г. Минск

Р4

АНАЛИЗ И СИНТЕЗ РЕЧИ

УДК 628.421

В.И.Александровский

ОСОБЕННОСТИ РЕАЛИЗАЦИИ МОДЕЛИ АНАЛИЗА, ОБНАРУЖЕНИЯ И ИНТЕРПРЕТАЦИИ СИГНАЛОВ СЛИТНОЙ РЕЧИ НА БАЗЕ ПЕРСОНАЛЬНОГО ВЫЧИСЛИТЕЛЬНОГО КОМПЛЕКСА

При отработке и исследовании различных методик распознавания речи и, в частности, при создании различных моделей анализа, обнаружения и интерпретации сигналов слитной речи одним из требований к реализации таких моделей на базе персонального вычислительного комплекса является необходимость уменьшения времени реакции модели до величин, приемлемых для человека. Именно в этом случае появляется возможность исследования особенностей модели на большом статистическом материале в приемлемые сроки.

В рассматриваемой модели анализа, обнаружения и интерпретации сигналов слитной речи, как и во всяком устройстве автоматической классификации, могут быть выделены следующие основные блоки: датчик, блок предварительной обработки, блок классификации, память эталонов [1].

Датчик, являющийся в нашем случае микрофоном, преобразует звуковые колебания в электрический сигнал, удобный для восприятия блоком предварительной обработки. Задача, решаемая блоком предварительной обработки, состоит в выделении характерных признаков речевого сигнала. Одной из важнейших операций на этом пути является выделение границ (хотя бы приблизительных) той области во входном звуковом сигнале, в которой находится анализируемая информационная часть (фраза, слово и т.п.). Помимо этого, блок предварительной обработки должен сформировать такой набор характерных признаков, которые были

25

бы удобна для эффективного разделения и классификации предъявляемых объектов - речевых сигналов. Блок классификации призван обеспечивать эффективное выявление з предъявлением речевом сигнале участков, с достаточной степенью достоверности соответствующих эталонам, находящимся в памяти эталонов, где последние были сформированы предварительно при обучении системы.

Блок предварительной обработки

Одной из основных задач блока предварительной обработки является выделение из входного потока информации той ее части, которая остается инвариантной к изменению мешающих факторов речевого и неречевого происхождения, к смене диктора и т.п.

Современный уровень развития микроэлектроники позволяет создать с использованием микросхем цифрового процессора обработки сигнала анализатор, фильтрующий многоканальное (до 30-40 каналов) спектральное признаковое описание речевого сигнала с периодом дискретизации 10 мс. Перед передачей признакового описания блоку классификации в блоке предварительной обработки должна быть произведена некоторая предварительная обработка описания, включающая:

- предварительную разметку сигнала на участки, отвечающие зонам "пауза" и "речь";
- выделение шумовой (неинформационной, неречевой) составляющей сигнала;
- компенсацию шумовой составляющей сигнала на участке "речевой" зоны;
- нормировку параметров и приведение их числовых значений к удобному для работы блока классификации диапазону.

Сочетание высокой плотности потока входной информации от анализатора спектра, достигающей 40-50 байт на периоде 10-30 мс, о необходимости ее обработки в реальном масштабе времени предъявляет специфические требования к операционному устройству блока предварительной обработки. Действительно, даже в предположении идентичности алгоритмов для каждого из, например, 50 байт входной информации, обновляющихся каждые 10 мс, получаем в среднем по 200 мкс на обработку одного байта, что позволяет в лучшем случае выполнить, применения распространенные микропро-

26

цессоры, около сотни операций, не считая ввода-вывода.

Однако существующие методики помимо действий, осуществляемых над всеми входными данными одинаково, предполагают ряд различных по характеру действий, выполняемых над выделенными группами данных. Так, в методике [2] помимо одинаковой для всех данных спектральных каналов цифровой фильтрации специализированной и неподобной друг на друга обработке подвергаются данные группы каналов низких и высоких частот, данные группы узкополосных и широкополосных каналов. Методика предполагает увязку результатов отдельных специализированных обработок между собой взаимозависимыми параметрами, а также слаживание за всем ходом процесса в целом, обеспечивающее формирование различных признаковых характеристик сигнала, выявление возможного начала и окончания полезного речевого сигнала и т.п.

Понятно, что такой и подобные ему алгоритмы предварительной обработки нельзя реализовать на однопроцессорной структуре, а создавать специализированные вычислители под отдельные фрагменты столь "многослойного" алгоритма, учитывая возможность модификации и развития последнего, экономически и технически не оправдано. Поэтому представляется целесообразным использовать при построении блока предварительной обработки конвейерную цепочку однотипных операционных блоков с возможностью внешней загрузки программного обеспечения.

Идентичность звеньев конвейера позволит легко при необходимости наращивать алгоритмическую мощность конвейера в целом, причем распределение алгоритмической нагрузки между процессорными звеньями конвейера осуществляется, исходя из возможности одного звена выполнять логически законченную часть обработки над всей порцией данных (или частью ее) за время до появления новой аналогичной порции данных.

Блок классификации

В основе алгоритма работы блока классификации лежит процесс выделения в предъявлении речевом сигнале $A = (a_1, a_2, \dots, a_i, \dots, a_n)$ отдельных участков-сегментов, которые "похожи" (по оговоренным критериям) на имеющиеся в распоряжении системы эталоны $B_k = (b_{k1}, b_{k2}, \dots, b_{ki}, \dots, b_{kj})$. Сам процесс

27

выделения очередного сегмента и "узнавания" в выделенном сегменте конкретного эталона основой на поэтапном сокращении исходного множества эталонов, участвующих в работе на текущем шаге [3].

На первом этапе гипотезируется начало сегмента с одновременным исключением из рассмотрения эталонов исходного множества, очевидно неподходящих по своим L_k первым отсчетам на гипотезируемое начало выделяемого сегмента. На втором этапе для каждого из оставшихся эталонов завершается гипотезация границ сегмента с возможным исключением нового числа эталонов, которые "неподходящи" уже по более широкому кругу признаков. Оставшиеся после второго этапа эталоны, снабженные числовыми оценками "степени подобия" (или "мерами доверия"), и составляют результат работы блока классификации. В этом смысле блок классификации выполняет роль "акустического процессора", являющегося базовым для следующих по иерархии уровням: семантического, pragmaticального и т.п. анализа.

Основные алгоритмические затраты как на первом, так и на втором этапах ложатся на организацию вычислений целевой функции

$$g_k(i,j) = \min \left\{ \begin{array}{l} g_k(i-1,j-1) + A_1 * d_k(i,j) \\ g_k(i-1,j) + A_2 * d_k(i,j) \\ g_k(i,j) + A_3 * d_k(i,j) \end{array} \right\}$$

обеспечивающей как нормализацию темпа речи, так и выработку оценок мер сходства эталонов и предъявляемой реализации. Целевая функция $g_k(i,j)$ имеет смысл расстояния в выбранном пространстве признаков между отрезком длиной i предъявляемой речевой реализации A и отрезком длиной j сопоставляемого эталона B_k .

Работа алгоритма по вычислению значений целевой функции $g_k(i,j)$ есть итерационный процесс, в котором для каждой итерации необходимо вычисление расстояния $d_k(i,j)$. Общее же число элементарных циклов как по вычислению значений локального расстояния $d_k(i,j)$, так и по вычислению значений целевой функции $g_k(i,j)$ зависит от числа эталонов K . Длины предъявляемой реализации I и длины участвующих в процессе эталонов G_k .

Исходя из двухэтапного характера работы алгоритма, общее число элементарных циклов вычислений N представлено суммой числа циклов первого этапа работы алгоритма N_1 и числа циклов второго этапа работы алгоритма N_2 :

$$N = N_1 + N_2.$$

Так как на первом этапе в обработке участвуют лишь первые L_k отсчетов эталонов, а число сопоставляемых эталонов может быть сужено за счет априорных сведений о семантических связях эталонов до величины $\beta_1 * k (\beta_1 < 1)$, то число элементарных циклов вычислений первого этапа составляет

$$N_1 = (\alpha_1 * I) * (\beta_1 * k) * L_k.$$

В приведенном выражении сомножитель $\alpha_1 * I$ показывает, что из всех отсчетов I предъявляемой реализации в вычислениях первого этапа участвует их часть $\alpha_1 < I$.

На втором этапе работы алгоритма участвуют уже все G_k отсчеты эталонов, но число самих эталонов сокращается по результатам работы первого этапа и составляет $\beta_2 * \beta_1 * k$ эталонов ($\beta_2 < 1$). С учетом того, что часть отсчетов предъявляемой реализации, не вошедшая в гипотезируемые сегменты, не будет участвовать в вычислениях, число элементарных циклов второго этапа составляет

$$N_2 = (\alpha_2 * I) * (\beta_1 * \beta_2 * k) * L_k.$$

В приведенном выражении $\alpha_2 < I$ определяет число отсчетов предъявляемой реализации, участвующих в вычислениях второго этапа алгоритма.

Исследования работы алгоритма показывают [4], что типичные значения коэффициентов α_1 и α_2 лежат в пределах 0,5 ... 0,7, а значение коэффициента β_2 обычно составляет 0,1 ... 0,3. Величина коэффициента β_1 определяется возможностями учесть семантические связи между элементами исходного словаря, а также той априорной информацией о предъявляемой реализации, которой может располагать акустический процессор до начала своей работы. В отсутствие такой информации необходимо полагать $\beta_1 = 1$.

Оцениваемые величины числа элементарных циклов вычислений d_k и g_k при условиях, типичных для большинства решаемых задач, лежат в пределах от $0,4 * 10^6$ до $4 * 10^6$. Конечно, помимо обес-

печения вычислений d_k и g_k на первом и втором этапах, требуется и другие работы: обеспечение задания начальных условий для обоих процессов, передача рабочих данных между процессами, систематизация рабочих данных и их анализ, разделение и оформление результатов и т.п. Однако быстроту реакции системы в конечном счете определяют оптимальность организации вычислений d_k и g_k как процессов, подавляющих все остальные по числу повторений.

Алгоритм вычисления локального расстояния $d_k(i,j)$ определяется выбором метрики – правила определения расстояния между двумя отсчетами в заданном пространстве признаков, выбранных для описания речевого сигнала. Метрика в виде суммы абсолютных значений разности значений одноименных признаков

$$d_k(i,j) = \sum_{p=1}^P |a_{pi} - b_{pj}|$$

является наиболее адекватной действительным различиям между реализациями речевых сигналов и она же наилучшим образом отвечает критерию простоты ее вычисления.

Даже для числа признаков $P = 4 \dots 8$ програмная реализация вычислений $d_k(i,j)$ для всех возможных элементарных циклов на любом из процессоров общего назначения составит время в десятки секунд, что не может быть признано приемлемым.

В таких условиях наиболее удачным решением блока классификации может быть аппаратная реализация вычислителя локальных расстояний и вычислителя целевой функции.

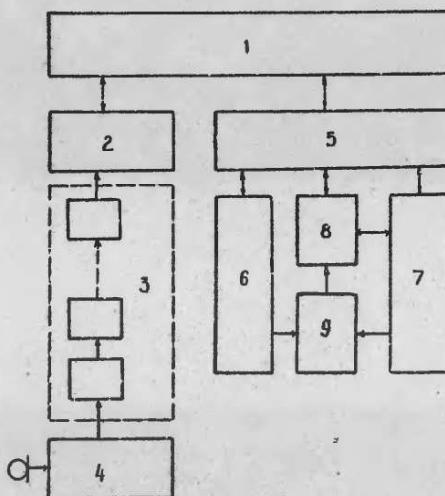
Конкретизация методики вычисления целевой функции и упорядоченность перебора индексов i и j позволит минимизировать число обращений к памяти рабочих данных.

Аппаратное разделение памяти рабочих данных (необходима для работы вычислителя целевой функции) и памяти эталонов (необходима для работы вычислителя локальных расстояний) при обеспечении их синхронного перебора позволит совместить по времени работу вычислителя локальных расстояний и вычислителя целевой функции. При этом время одного рабочего цикла подобного вычислителя будет определяться структурой памяти эталонов и рабочих данных, определяющей, в свою очередь, число обращений к памяти для записи или чтения данных, относящихся к одному отсчету.

30

Архитектура аппаратной реализации модели

Учитывая вышеизложенное, при реализации модели анализа, обнаружения и интерпретации сигналов слитной речи на базе персонального вычислительного комплекса может быть предложена структурная схема, изображенная на рисунке.



Структурная схема реализации модели анализа, обнаружения и интерпретации сигналов слитной речи на базе персонального вычислительного комплекса

К ПЭВМ 1 через интерфейсный блок 2 подключен конвейер процессоров предварительной обработки 3, принимающий и обрабатывающий сигналы с блока спектроанализатора 4, а через другой интерфейсный блок 5 – синхронно адресуемое память эталонов б

и память языческих данных 7, выходные данные которых используются в аппаратных вычислителях локального расстояния 8 и целевой функции 9.

Реализация модели анализа, обнаружения и интерпретации сигналов слитной речи в соответствии с предлагаемой архитектурой позволит не только проводить разнообразные исследования особенностей модели на большом статистическом материале, но и послужит основой для разработки устройства речевого ввода, работающего в реальном масштабе времени.

ЛИТЕРАТУРА

1. Фор А. Восприятие и распознавание образов. - М., 1989.
2. Дегтярев Н.П. Акустическое описание артикуляции параметрами обобщенных формант спектра речи // Автоматическое распознавание слуховых образов (APCO-I5). - Таллинн, 1999. - С. 145.
3. Дегтярев Н.П. Алгоритм распознавания слов в непрерывном сигнале // Бионика интеллекта: Мат. докл. Всесоюз. симпозиума "Бионика интеллекта". - Харьков, 1987. - С.27.
4. Дегтярев Н.П., Александровский В.И., Александровская М.Н. Исследования алгоритма обнаружения и распознавания слов слитной речи // Автоматическое распознавание слуховых образов (APCO-I5). - Таллинн, 1999. - С. 93.

г. Минск

32

АНАЛИЗ И СИНТЕЗ РЕЧИ

УДК 007.001.362 + 681.327

А.С.Рылов

ИССЛЕДОВАНИЕ МЕТОДА ОЦЕНКИ ДЛИНЫ РЕЧЕОБРАЗУЮЩЕГО ТРАКТА ПО КОЭФФИЦИЕНТАМ ПОГЛОЩЕНИЯ

Длина речеобразующего тракта L является важным геометрическим параметром, от которого в значительной степени зависят акустические параметры. Например, L обратно пропорциональна средним значениям формантных частот F_n . А как известно [1], F_n для одной и той же гласной фонемы, произнесенной средним женским голосом, оказывается на 20% выше соответствующих значений F_n для среднего мужского голоса и на 20% ниже F_n для той же гласной, произнесенной восьмилетним ребенком.

Таким образом, L может быть критерием качества в системах распознавания речи и должна непрерывно определяться наряду с другими акустическими параметрами, которые в зависимости от нее должны непрерывно корректироваться.

Предлагаемый метод расчета длины речеобразующего тракта базируется на модели речеобразования, согласно которой речевой тракт человека от связок до губ представляется в виде многосекционной трубы. Секции трубы имеют одинаковую длину, но разные диаметры сечений. Такая труба аппроксимируется последовательно соединенными симметричными четырехполюсниками [2]. Каждый из них имеет свою передаточную функцию

$$H_i(z) = P_i(z) / P_{i+1}(z), \quad (1)$$

где P - давление воздуха в i -й секции.

Передаточная функция всей системы определяется произведением передаточных функций отдельных четырехполюсников [3]:

Зак. №

33

$$H(x) = H_1(x) \cdot H_2(x) \cdots H_M(x) = \prod_{i=1}^M H_i(x) = P_1(x)/P_M(x).$$

Если труба разомкнута в губах и почти закрыта у источника (см. рисунок), то в момент $t=0$ на ее вход поступает импульс воздушного потока $P^+(l)$, который, распространяясь со скоростью звука, через время $t=l/c$ достигнет выхода и, отразившись, начнет двигаться назад к источнику. Таким образом, на выходе будет разность падающей и отраженной волн. Отношение амплитуд отраженной и падающей волн называется коэффициентом отражения [2], т.е.

$$\kappa_p = P^-(l) / P^+(l). \quad (3)$$

Из выражения для (3) легко получить коэффициент поглощения звука импедансом в трубе:

$$\alpha = 1 - \kappa_p^2 = \frac{E_n - E_r}{E_n} = 1 - \frac{E_r}{E_n}, \quad (4)$$

где E_n - интенсивность падающей волны,

E_r - интенсивность отраженной волны, а разность $E_n - E_r$ представляет собой энергию, поглощенную импедансом Z_l на выходе из трубы.

Выражение (4) справедливо, так как

$$E_r/E_n = \{P^-(l)/P^+(l)\}^2.$$

По аналогии с (4) определяется коэффициент поглощения для многосекционной трубы.

Так как $E_{n,i} = P_i^2$, то $E_{n,i} - E_{r,i} = E_{n,i-1} \cdot P_{i-1}^2$, а $\alpha_i = P_i^2 / P_{i-1}^2$.

Последняя формула тождественна квадрату передаточной функции (I) для одной секции трубы. Для многосекционной трубы по аналогии с (2) получается

$$\alpha = P_n^2 / P_M^2 \prod_{i=1}^M (1 - \kappa_i^2). \quad (5)$$

Коэффициент поглощения речевого тракта α может быть оценен непосредственно из речевого сигнала. Необходимо лишь определить энергию поглощения речевого тракта ψ . Для этого сформулируем два утверждения:

Утверждение I. Энергия звуковой волны в результате много-

кратного прохождения ее через одну и ту же трубу уменьшается каждый раз на одну и ту же величину ($\psi = \cos \alpha t$).

Утверждение 2. Если падающая волна речевого сигнала, представляющая собой множество $X^+ \{x_n^+ | n=1,2,\dots,N\}$ дискретных значений, пропускается через авторегressiveйный фильтр, коэффициенты авторегрессии или коэффициенты отражения которого были выделены из того же множества X^+ , то на выходе фильтра синтезируется множество $X^- \{x_n^- | n=1,2,\dots,N\}$, представляющее собой отраженную звуковую волну речевого сигнала, т.е.

$$X^- = \sum_{i=1}^M \alpha_i x^{+(n-i)}. \quad (6)$$

Если теперь принять сигнал на входе речевого тракта за падающую волну X^+ , а сигнал X^- в (6) за отраженную, то разность энергий этих сигналов даст энергию поглощения речевого тракта ψ , которая с учетом сформулированных утверждений записывается в следующем виде:

$$\psi = \sum_{n=1}^N (x_n^+)^2 - \sum_{n=1}^N \left(\sum_{i=1}^M \alpha_i x^{+(n-i)} \right)^2 = E_n - E_r. \quad (7)$$

В формулах (6), (7) α_i - параметры авторегрессии, которые вместе с коэффициентами отражения определяются с помощью рекуррентного алгоритма Левинсона [4] (8) - (II):

$$\alpha_{m+1,i} = \begin{cases} \alpha_{m,i} & i=0 \\ \alpha_{m,i} + \alpha_{m,m+1-i} k_m & i=1,2,\dots,m \\ k_m & i=m+1 \end{cases} \quad (8)$$

$$\alpha_{m+1} = \sum_{i=0}^{m+1} \alpha_{m+1,i} r_i; \quad (9)$$

$$\beta_m = \sum_{i=0}^m \alpha_{m+1,i} r_{m+2-i}; \quad (10)$$

$$k_m = \beta_m / \alpha_{m+1}. \quad (II)$$

Начальные условия: $\alpha_{00}=1; \alpha_0=r_0; \beta_0=r_1$, где r_i - коэффициенты корреляции входного речевого сигнала.

Используя (6), (7) и (4), начиная с i -й рекурсии в процес-

дуре Левинсона, определяем коэффициенты поглощения $\alpha_i, \alpha_{i+1}, \dots, \alpha_j, \dots, \alpha_M$.

M – соответствует самому большому возможному значению L в формуле [4]:

$$M = 2f_d L, \quad (12)$$

а i – самому минимальному значению L в (12). Здесь c – скорость звука и f_d – частота дискретизации, которые являются константами. Задав таким образом множество $\alpha_i, \alpha_{i+1}, \dots, \alpha_j, \dots, \alpha_M$, определяем максимум 1-й производной функции $\alpha(M)$. Значение j ($i \leq j \leq M$, для которого найден $\max(\alpha_j - \alpha_{j-1})$) подставляется в (12) и определяется L .

В табл. I, являющейся распечаткой промежуточных результатов работы программы расчета длины речевого тракта, приведены значения коэффициентов поглощения $\alpha = ER/EP$ для 10 кадров звука /И/ длительностью по 20 мс каждый. Причем $i=8$, $M=20$, однако для каждого кадра распечатано также 7-е значение ER/EP , так как

$$\Delta\alpha_8 = \alpha_8 - \alpha_7.$$

В табл. I взяты в рамочку те значения α в каждом кадре, которые дают $\Delta\alpha_j$ – максимальные значения. Например, для второго кадра значение M , для которого $\Delta\alpha_j$ максимально, равно 15, т.к. разность $\alpha_{15} - \alpha_{14} = 0,89267I - 0,863790 = 0,02888I$ является наибольшей.

Если посмотреть на остальные кадры, то $\Delta\alpha_{\max}$ будет почти всегда при $M=15$. Тем не менее в кадрах 3, 4, 6, 9 в рамку взяты значения α_{16} . Дело в том, что после нахождения основного максимума $\Delta\alpha_{\max}$ бывают максимумы того же порядка $\Delta\alpha_{j+m}$, расположенные на промежутке от j до M .

Достоверную оценку L дают значения j , соответствующие именно такому последнему максимуму. На рис. 1 показан график функции $\Delta\alpha(M)$ для 3-го кадра. Все значения $\Delta\alpha_8^{(20)}$ уложены на 100.

В табл. 2 помещены результаты расчета L и значения j (обозначены буквой M) для упомянутых 10 кадров фонемы /И/.

На рис. 2 и рис. 3 приведены форма сигнала и функция площади сечений соответственно для первого кадра фонемы /И/.

Таблица I

Значения коэффициентов поглощения для звука /И/					
ER/EP = 0.803761	ER/EP = 0.805278	ER/EP = 0.807442	ER/EP = 0.805982		
ER/EP = 0.832827	ER/EP = 0.846306	ER/EP = 0.846306	ER/EP = 0.850484		
ER/EP = 0.874322	ER/EP = 0.876220	ER/EP = 0.879434	ER/EP = 0.871982		
ER/EP = 0.880167	ER/EP = 0.881687				
ER/EP = 0.8199915	ER/EP = 0.822349	ER/EP = 0.825429	ER/EP = 0.839007		
ER/EP = 0.844083	ER/EP = 0.855469	ER/EP = 0.862013	ER/EP = 0.863790		
ER/EP = 0.892671	ER/EP = 0.897213	ER/EP = 0.900195	ER/EP = 0.900196		
ER/EP = 0.900909	ER/EP = 0.901559				
ER/EP = 0.803985	ER/EP = 0.803998	ER/EP = 0.804115	ER/EP = 0.811791		
ER/EP = 0.842319	ER/EP = 0.850829	ER/EP = 0.856310	ER/EP = 0.859796		
ER/EP = 0.862781	ER/EP = 0.893238	ER/EP = 0.902279	ER/EP = 0.902740		
ER/EP = 0.902865	ER/EP = 0.903361				
ER/EP = 0.864018	ER/EP = 0.864800	ER/EP = 0.864328	ER/EP = 0.837186		
ER/EP = 0.849016	ER/EP = 0.854917	ER/EP = 0.862613	ER/EP = 0.865414		
ER/EP = 0.892714	ER/EP = 0.901053	ER/EP = 0.906705	ER/EP = 0.907146		
ER/EP = 0.907235	ER/EP = 0.907976				
ER/EP = 0.798848	ER/EP = 0.798987	ER/EP = 0.799118	ER/EP = 0.818490		
ER/EP = 0.825058	ER/EP = 0.835522	ER/EP = 0.845611	ER/EP = 0.847743		
ER/EP = 0.878766	ER/EP = 0.884888	ER/EP = 0.887596	ER/EP = 0.887625		
ER/EP = 0.887833	ER/EP = 0.888599				

Таблица 2

Значения *L* для фонемы /И/

```

ER/EP = 0.782008 ER/EP = 0.782293 ER/EP = 0.782655 ER/EP = 0.798369
ER/EP = 0.603810 ER/EP = 0.810105 ER/EP = 0.826400 ER/EP = 0.829550
ER/EP = 0.848405 ER/EP = 0.856610 ER/EP = 0.859163 ER/EP = 0.859398
ER/EP = 0.859777 ER/EP = 0.862290
ER/EP = 0.805891 ER/EP = 0.805910 ER/EP = 0.806022 ER/EP = 0.831742
ER/EP = 0.843569 ER/EP = 0.850456 ER/EP = 0.861525 ER/EP = 0.862867
[ER/EP = 0.897135] ER/EP = 0.903292 ER/EP = 0.909874 ER/EP = 0.910285
ER/EP = 0.910233 ER/EP = 0.910707
ER/EP = 0.809752 ER/EP = 0.810318 ER/EP = 0.810332 ER/EP = 0.828021
ER/EP = 0.834418 ER/EP = 0.845066 ER/EP = 0.852951 ER/EP = 0.853921
[ER/EP = 0.888351] ER/EP = 0.892923 ER/EP = 0.895942 ER/EP = 0.895985
ER/EP = 0.897082 ER/EP = 0.897473
ER/EP = 0.790779 ER/EP = 0.791673 ER/EP = 0.791707 ER/EP = 0.810082
ER/EP = 0.818234 ER/EP = 0.823721 ER/EP = 0.830236 ER/EP = 0.836651
ER/EP = 0.863809 ER/EP = 0.873265 ER/EP = 0.875198 ER/EP = 0.875127
ER/EP = 0.876529 ER/EP = 0.877661
ER/EP = 0.809109 ER/EP = 0.810779 ER/EP = 0.811999 ER/EP = 0.827985
ER/EP = 0.830095 ER/EP = 0.842731 ER/EP = 0.847355 ER/EP = 0.849421
[ER/EP = 0.885044] ER/EP = 0.839194 ER/EP = 0.893558 ER/EP = 0.893558
ER/EP = 0.896042 ER/EP = 0.896052

```

anout.txt

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 17,0 M = 16

***** Выполнение *****

L = 17,0 M = 16

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 17,0 M = 16

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 17,0 M = 16

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

L = 15,9 M = 15

***** Выполнение *****

Таблица 3

Значения *L* для звуков /И/, /А/, /О/

Фонемы	Ф	П3	Р
И	16,5	14,6	15,9
А	17,0	16,4	17,0
О	18,5	17,7	18,1

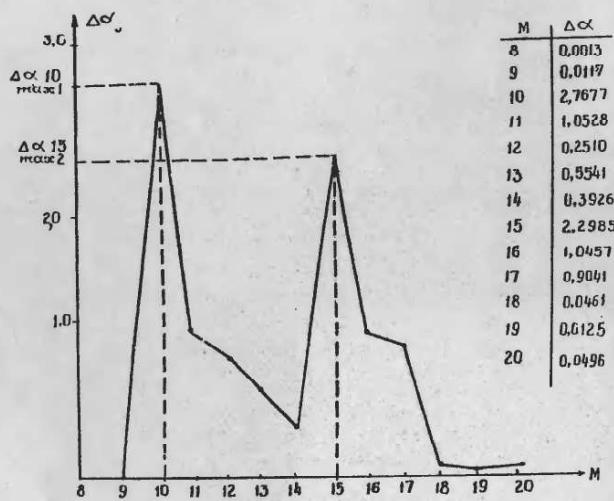


Рис. 1. График функции $\Delta\alpha(M)$

Комментируя эти рисунки, следует отметить, что ФПС на рис. 2, б рассчитывалась для $M=20$. Однако, как видно из рисунков, на самом деле $M=15$. В табл. 2 $j=M=15$ – именно то значение, после которого ФПС на рис. 2, б практически равна нулю.

На рис. 2, б рядом с рассчитанной ФПС показана конфигурация ФПС по Фанту [5]. Значения I_1 , рассчитанные вышеописанным методом для фонем /И/, /А/, /О/, соотвествуют значениям I_1 , приведенным Фантом, и несколько отличаются от значений Пейджа и Зу [6]. В табл. 3 приведены эти значения. Под буквой Φ – значение по Фанту, под буквами П.З – значения по Пейджу и Зу, под буквой Р – по Рылову.

40

M	$\Delta\alpha$
8	0,0013
9	0,0117
10	2,7677
11	1,0528
12	0,2510
13	0,5541
14	0,3926
15	2,2985
16	1,0457
17	0,9041
18	0,0461
19	0,0125
20	0,0496

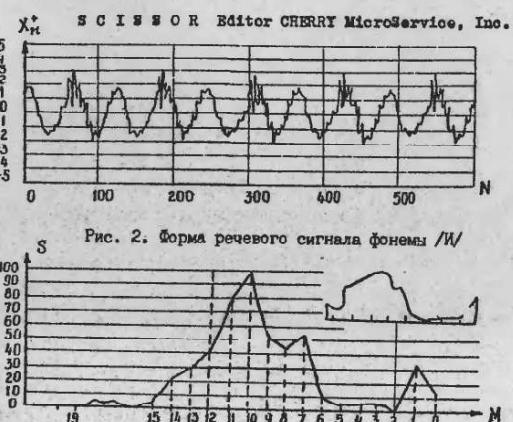


Рис. 2. Форма речевого сигнала фонемы /И/

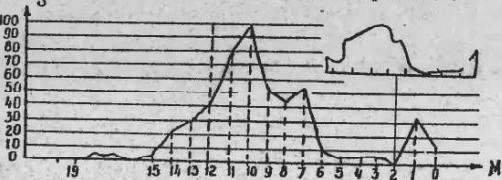


Рис. 3. Функция площадей сечения фонемы /И/

ЛИТЕРАТУРА

- Фант Г. Анализ и синтез речи. – Новосибирск: Наука, 1970.
 Ржевкин С.Н. Курс лекций по теории звука. – М.: МГУ, 1960.
 Джури Э. Импульсные системы автоматического регулирования. – М.: Госиздат. Физ.-матем. лит., 1963.
 Маркел Д.Д., Грей А.Х. Линейное предсказание речи. – М.; Связь, 1990.
 Фант Г. Акустическая теория речеобразования. – М.: Наука, 1964.
 Paige A, Zue V. Calculation of vocal Tract Length IEEE Transactions on Audio and Electroacoustics 1970.– Vol. AU – 18. – №3.

41

Минск

АНАЛИЗ И СИНТЕЗ РЕЧИ

УДК 007.001.326 + 691.327

А.С.Рылов, Т.В.Левковская

ЧАСТОТНО-АДАПТИВНЫЙ АВТОРЕГРЕССИОННЫЙ АНАЛИЗ РЕЧЕВЫХ СИГНАЛОВ

Введение

Акустический анализ речи является низшим уровнем иерархии в системах распознавания речевых образов. Его результаты служат исходным материалом для лексического, синтаксического и семантического анализов. Естественно, что ошибки, сделанные на акустическом уровне, имеют большее значение, чем ошибки, возникающие позже. Вся информация, которая отбрасывается на уровне акустического анализа, становится недоступной для остальной части системы. Поэтому одной из главных причин столь неенакительных успехов в области распознавания речи на фоне успехов технологий вычислительной техники является, на наш взгляд, проблема первичного описания речевых сигналов. Трудности, которые здесь возникают, связаны с преодолением априорной неопределенности речи.

Эффективным способом преодоления априорной неопределенности при решении задач приема и обработки информации является создание адаптивных систем. При этом под адаптацией понимается обучение, самообучение, а также процесс оптимальной перестройки структуры приемного устройства в соответствии с критерием качества. Простым примером автомата для автоматической адаптивной системы является автоматическая регулировка усиления (APU), изменяющая в радио- и телевизионных приемниках. Функция этой системы - уменьшение чувствительности приемника при увеличении среднего уровня входного сигнала. Таким образом, приемник может адаптироваться к широкому диапазону уровней входных сигналов.

и формировать значительно более узкий диапазон уровней выходных сигналов.

В данной работе предпринята попытка создать адаптивный анализатор, позволяющий снизить априорную неопределенность речевого сигнала за счет предварительной оценки геометрии органов артикуляции речесформирующих трактов диктором и адаптировать широкий диапазон входных сигналов к более узкому.

Обоснование выбора метода первичного анализа речи

В настоящее время получили широкое распространение два метода первичного анализа речи. Это спектрально-полосный и авторегрессионный.

Суть спектрально-полосного анализа заключается в фильтрации речевого сигнала гребенкой полосовых фильтров с последующим детектированием и НЧ-фильтрацией в каждом канале. В конечном итоге с выхода каждого канала снимается информация о распределении энергии в каждой частотной полосе.

Авторегрессионный метод базируется на методе наименьших квадратов, предложенном Гауссом еще в 1875 г. и получившем название "метода линейного предсказания" в работах Винера. Дальнейшее развитие для анализа речи этот метод получил в работах Киприева [1,2], Сaito и Итакуры [3], Атала и Шредера [4], Мартина и Грэя [5] и др.

Говоря о спектрально-полосном методе, следует отметить два его важнейших достоинства - простоту и полнокомпактность. Позитиву в настоящее время в приборах для акустических исследований речи и при разработке распознавающих устройств с настройкой диктора применяется именно спектрально-полосный принцип анализа.

Однако возникают и существенные трудности при анализе речи этим способом. Это связано с тем, что если F_0 (частота колебаний голосовых связок) имеет частоту, сравнимую с полосой просканивания канальных фильтров или превышающую ее, то на динамических спектрограммах (спектрограммах) проявляется периодичность излучения в виде тонкой структуры гармоник по оси частот, отличающихся друг от друга на величину F_0 . В результате получается довольно запутанная картина, которая не позволяет следить за

43

движением формант.

Кроме того, если говорить о спектральных срезах, которые используются в аппаратуре распознавания речи, то, как известно, максимумы спектра характеризуют форманты, но если частота f_0 больше, расстояние между гармониками велико, и тогда возникает ошибка в определении частот формант по гармоническому спектру на величину порядка $f_0/4$ [6]. Поэтому с критичностью соотносим ширину полосы фильтров анализатора и частоты основного тона f_0 , связан тот факт, что спектрально-полосный анализ лучше подходит для мужских голосов, чем для женских.

При анализе речи авторегрессионным методом влияние частоты основного тона на качество анализа практически отсутствует. Поэтому для первичного анализа речевых сигналов на акустической уровне целесообразно использовать авторегрессионный анализ как анализ, менее подверженный влиянию функции источника на функции тракта.

Теоретическое обоснование принципа аддитивного авторегрессионного анализа

Говоря о преимуществах авторегрессионного анализа, необходимо помнить, что для некоторых голосов и типов сегментов формантные частоты могут не присутствовать или же, напротив, могут появляться слишком большие пики в спектре, похожие на форманты [7]. Главной причиной таких спектральных аномалий, на взгляд, является то, что при анализе не выполняется одно из важных условий авторегрессионного анализа речи [5], согласно которому $f_{cr} = f_0/2$, где f_{cr} - частота среза ФНЧ, f_0 - частота дискретизации АЦП.

Действительно, в случае, когда $f_{cr} > f_0/2$, возникает известный эффект наложения спектров, а при $f_{cr} < f_0/2$ увеличивается объем памяти, вычислений и, самое главное, ухудшается качество процесса "обеления" авторегрессионного анализа. Это доказывается в [5] с помощью введенной меры равномерности спектров

$$C(F) = \exp \left[\int_{-\infty}^{\infty} V(\theta) \frac{d\theta}{2\pi} \right],$$

которая является количественной оценкой эффективности ступени обеления и представляет собой усредненное значение нормированного логарифмического спектра,

$$\text{где } V(\theta) = \ln [|F(e^{j\theta})|^2 / r_f(0)| -$$

нормированный логарифмический спектр, который в идеальном случае равен 0. $r_f(0) = \sum_{n=0}^{\infty} f^2(n)$ - энергия сигнала. Действительно, если перед АЦП производится фильтрация о частотной срезе, которая значительно меньше $f_0/2$, то дискретный сигнал будет иметь малые значения спектра для части частотного диапазона. Это приведет к большим отрицательным значениям нормированного логарифмического спектра $V(\theta)$, а это, в свою очередь, к уменьшению спектральной равномерности $C(F)$.

С другой стороны, известно [5], что для адекватного представления голосового тракта в идеальных условиях память модели $A(Z)$ должна быть такой, чтобы обеспечить хранение сигнала на интервале, равном $2L/C$, где L - длина речевого тракта, C - скорость звука (34000 см/с), M - порядок авторегрессии или количество отрезков труб длины L , с помощью которых аппроксимируется речевой тракт.

Для выбранной модели период дискретизации определяется $T = 1/f_0 = 2L/C$. В результате получается соотношение

$$M/f_0 = 2L/c. \quad (I)$$

Если в (I) M и c константы, то f_0 и L должны быть переменными, но изменение L предполагает наличие функции $f_{cr}(L)$. Такая функция изображена на рис. I. Она получена с помощью (I) для $M = 8$ и $c = 34000$ см/с. Действительно, полная длина речеобразующего тракта обратно пропорциональна средним значениям его формант [6]. Если разные дикторы произнесут одну и ту же гласную, то соответствующая относительная разница в длинах их речеобразующих трактов будет перенесена обратно пропорционально с тем же коэффициентом на частотный спектр. Слушатель же эту разницу практически игнорирует.

Однако простые схемы подстройки под диктора при помощи линейного масштабирования формантных частот ограничиваются сильно меняющимися различиями масштабных множителей для отдельных

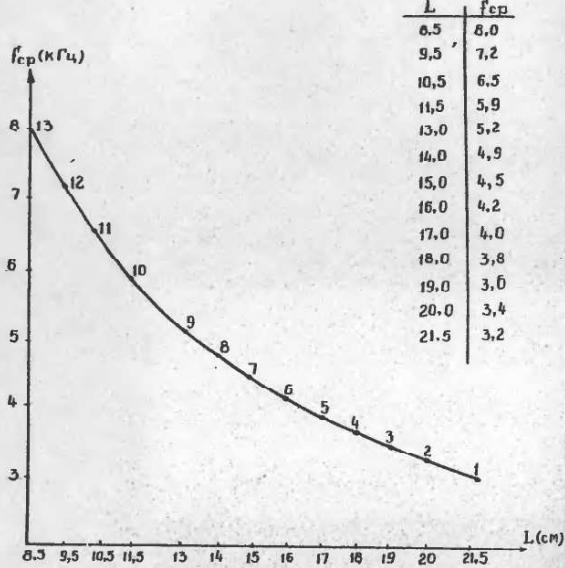


Рис. I. Функция зависимости $f_{cp}(L)$

формант у различных звуков [6]. По-видимому, это происходит по двум причинам: во-первых, длина речевого тракта сильно изменяется при произношении различных фонем одним и тем же диктором (например, для гласных до 13% [8]); во-вторых, специалистам хорошо известно, что результаты анализа речевых сигналов на периоде смыкания голосовых связок и на периоде их размыкания существенно отличаются. Это происходит потому, что при размыкании связок увеличивается длина акустического резонатора за счет подключения трахеи. Поэтому следует ожидать, что на периоде размыкания голосовых связок значение L может резко увеличиться.

46

Все эти рассуждения наводят на мысль о том, что существует необходимость постоянного измерения L в потоке речи и регулирования в соответствии с ней частотного диапазона входного речевого сигнала, адаптируя при этом параметры авторегрессионного фильтра. $f_{cp}(L)$ должна быть априорной функцией по отношению к авторегрессионному процессу.

Структура алгоритмов функционирования частотноадаптивного авторегрессионного спектрального анализа речи

Структура алгоритмов частотноадаптивного авторегрессионного анализа параметров изображена на рис.2.

В ее состав входят:

алгоритм определения длины речевого тракта - DLRT;

фильтры низкой частоты - FILTR;

алгоритм прореживания - INTERP-0;

алгоритмы функционирования авторегрессионного анализа - LPCAN;

алгоритм расчета спектральных срезов - SPEKTR.

Порядок их функционирования следующий. Из входной последовательности определяется L - длина речевого тракта. Затем в зависимости от L выбирается тот или иной фильтр. Всего фильтров 9 (в соответствии с рис.2). Так как частота дискретизации после каждого фильтра должна быть одна, то после фильтрации работает алгоритм прореживания. Далее работает алгоритм авторегрессионного анализа, на выходе которого формируется массив параметров авторегрессии, спектральных срезов и построения динамических спектрограмм. Рассмотрим работу этих алгоритмов.

Алгоритм расчета длины речевого тракта (DLRT) описан в данном сборнике.

Расчет и структура ФНЧ

В качестве ФНЧ использовались каскадные эллиптические фильтры, состоящие из звеньев 1-го и 2-го порядков (рис.3). Методику их расчета и необходимые нормограммы можно найти в [9]. Как уже упоминалось, используется 9 фильтров (табл.1).

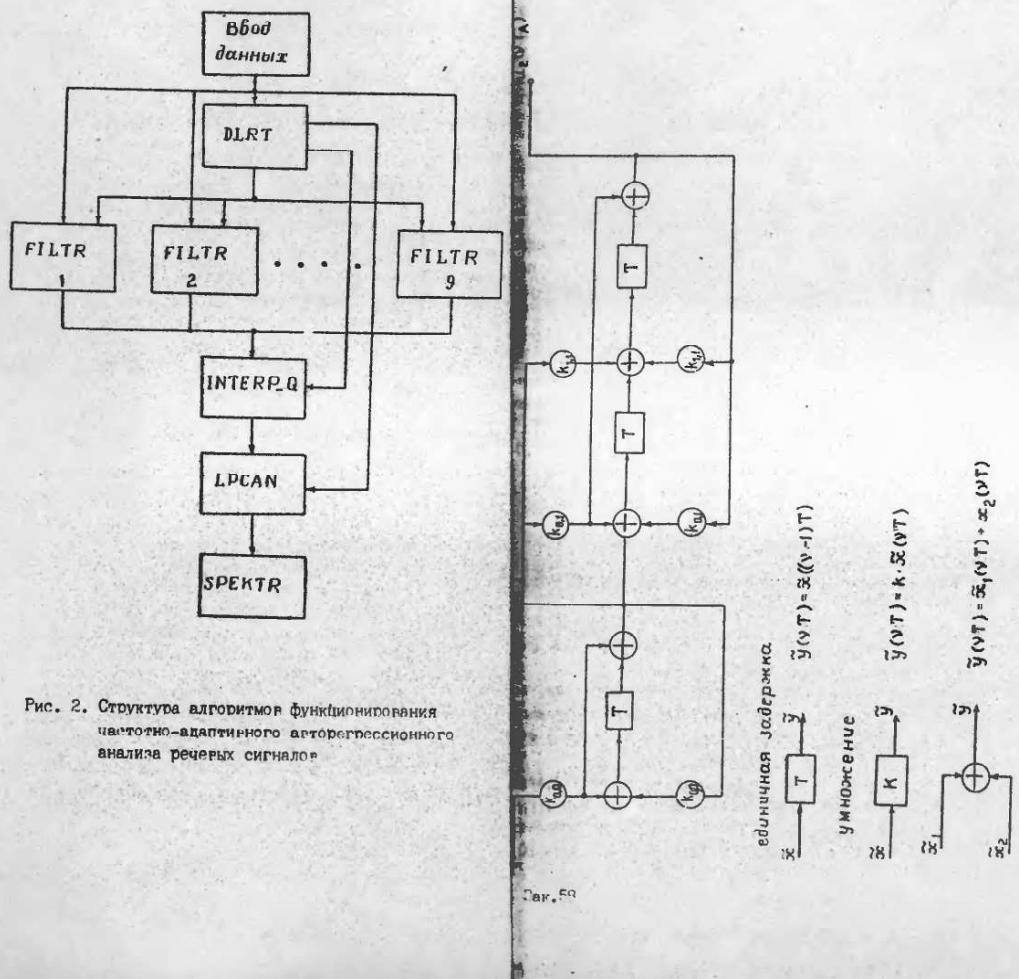


Таблица I

# п/п	1	2	3	4	5	6	7	8	9
$L, \text{см}$	23,80	22,10	20,40	19,70	17,00	15,30	13,60	11,90	10,20
$f_{\text{ср}} \text{Гц}$	2,14	2,31	2,50	2,73	3,00	3,33	3,75	4,29	5,00

Параметры одного из фильтров и амплитудно-частотная характеристика (АЧХ) приведены на рис.4. Как видно из рисунка, фильтр имеет крутой спад АЧХ после $f_{\text{ср}}$. Минимальное затухание в полосе задерживания не менее 40 дБ, а максимальное затухание в полосе пропускания - не более 0,1773 дБ. Фильтры 1-5 имеют 7-й порядок, фильтры 6-9 - 5-й порядок.

Алгоритм прореживания (INTERP_0)

Если f_{d_1} - базовая частота дискретизации и $f(nT_1)$ - значение отсчета в момент nT_1 , для данной f_{d_1} , а f_{d_2} - частота дискретизации, для которой необходимо найти значение $f(mT_2)$ в момент $mT_2 < nT_2 < (n+1)T_1$, то алгоритм определения $f(mT_2)$ сводится к следующему:

Сначала проверяются условия:

1. Если $nT_1 < mT_2 < (n+1)T_1$, то $n = n+1$.
2. Если $nT_1 = mT_2$, то $f(mT_2) = f(nT_1)$.
3. Если $(n+1)T_1 = mT_2$, то $f(mT_2) = f((n+1)T_1)$.
4. Если $nT_1 < mT_2 < (n+1)T_1$ и $f(nT_1) = f((n+1)T_1)$, то $f(mT_2) = f(nT_1) - f((n+1)T_1)$.
5. Если $nT_1 < mT_2 < (n+1)T_1$ и $f(nT_1) \neq f((n+1)T_1)$, то
1) $f((n+1)T_1) - f(nT_1) = a$; 2) $(mT_2 - nT_1)T_1 = c$; 3) $I = c/a$
4) $b = d - a$; 5) $f(mT_2) = f((n+1)T_1) - b$.

Рис. 5 поясняет принцип работы алгоритма прореживания, использованного для преобразования дискретных значений речевого сигнала при переходе от f_{d_1} к f_{d_2} . Причем надо помнить, что $f_{d_1} > f_{d_2}$.

Для правильной работы данного алгоритма необходимо все же знать, между какими значениями n и $n+1$ находится $m = (n+1)T_1$.

На рис.6 изображены графики двух кривых: а - это исходная последовательность речевого сигнала, имеющая длительность 2

```

0 [0]=0.215874, k0[0]=0.568253
0 [1]=0.550312, k0[1]=-0.537525, k1[1]=-0.026370, k1[1]=1.116394
0 [2]=-0.204254, k0[2]=-0.818050, k1[2]=-0.205783, k1[2]=-1.090342
0 [3]=-0.066213, k0[3]=-0.944451, k1[3]=-0.069542, k2[3]=1.079636
0 [4]=15.031651,k0[4]=-0.988183, k1[4]=-12.482254,k3[4]=1.079747
  \LPC>

```

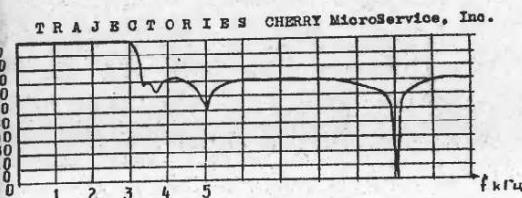


Рис. 4. АЧХ фильтра N 5

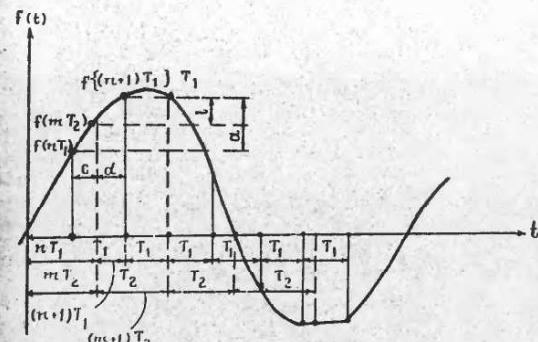


Рис. 5. График, поясняющий принцип работы алгоритма прореживания

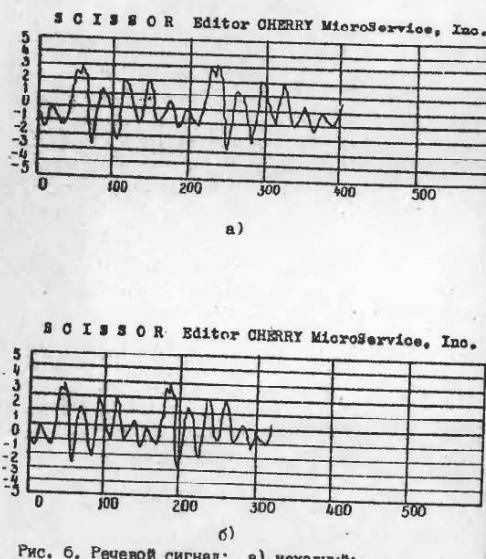


Рис. 6. Речевой сигнал: а) исходный;
б) после прореживания

52

(400 отсчетов) и $f_{d_1} = 20 \text{ кГц}$ или $T_1 = 1/f_{d_1} = 50 \text{ мкс}$; б - это тот же речевой сигнал, но уже с $f_{d_2} = 16 \text{ кГц}$ или $T_2 = 1/f_{d_2} = 63 \text{ мкс}$ (320 отсчетов).

Алгоритмы авторегрессионного анализа (LPCAN)

В состав LPCAN входят два алгоритма - RREEM и KPRMC. RREEM выполняет предварительную обработку сигнала. Если $x(n)$ - отсчеты речевого сигнала, а N - количество отсчетов в кадре, то на выходе RREEM в зависимости от параметра L (вариант работы RREEM) будем иметь

$$L=1 \quad x'(n) = w(n)[x(n) - \mu x(n-1)]; \quad \mu = R_1/R_0; \quad R_0 = \sum_{n=1}^N x(n)^2$$

$$w(n) = 0,54 - 0,46 \cos \frac{2\pi n}{N} \quad R_1 = \sum_{n=1}^N x(n)x(n-1)$$

$$L=2 \quad x'(n) = w(n)x(n)$$

$$L=3 \quad x'(n) = w(n)[x(n) - x(n-1)]$$

$$L=4 \quad x'(n) = x(n).$$

KPRMC вычисляет α -и k -параметры, используя описанную в [5] процедуру Левинсона.

Алгоритм расчета спектральных срезов (SPEKTR)

Для расчета спектральных срезов по параметрам авторегрессии определяется модуль передаточной функции авторегрессионного фильтра синтеза

$$H(z) = 1 / \sum_{i=0}^M \alpha_i z^{-i},$$

где $z = \exp(i2\pi n/N) = \cos i2\pi n/N - j \sin i2\pi n/N$;

α_i - параметры авторегрессии, N - число; характеризующее разрешающую способность спектра, $n = 1, 2, \dots, N/2$.

Таким образом, логарифмический спектр рассчитывается по формуле

$$S = -20 \lg \left| \sum_{i=0}^M \alpha_i \cos i2\pi n/N - j \sum_{i=0}^M \alpha_i \sin i2\pi n/N \right|.$$

Результаты экспериментальных исследований

Проверка алгоритма частотно-адаптивного авторегрессионно-

53

го анализа осуществлялась на стационарных фрагментах отдельно произносимых гласных фонем.

Речевые сигналы вводились в ЭВМ с помощью 12'-разрядного АЦП после предварительной обработки в аналоговом блоке автоматической регулировки входных уровней речевого сигнала. $f_{ср}$ в ФНЧ была выбрана равной 5 кГц, а f_d - в АЦП 10 кГц. Затем сигналы обрабатывались в описанном выше частотно-адаптивном авторегрессионном анализаторе, на выходе которого получались спектральные срезы, рассчитанные с помощью (2). Для анализа использовались следующие параметры: $L=3$ (в RREEM); $M=6$ в (КРМС); разрешающая способность спектра $\Delta S = f_d/N = 10000/200 = 50$ Гц.

На рис. 7 и 8 представлены спектральные срезы фонем /O/ и /И/. На рис. 7, а и 8, а изображены срезы, полученные не адаптивным авторегрессионным анализатором, ч) на рис. 7, б и 8, б - срезы, полученные описанным выше анализатором. Вторая форманта фонемы /O/ на рис. 7, б не имеет ярко выраженного пика, в то время как на рис. 7, б вторая и третья форманты выделены четко. На рис. 8, а в области 3 кГц имеется ложная форманта, которая исчезла при анализе частотно-адаптивным авторегрессионным анализатором рис. 8, б.

Аналогичные результаты получались и для других фонем, хотя были и редкие случаи, когда не удавалось разделить две очень близко расположенные форманты.

Оценка работы частотно-адаптивного авторегрессионного анализатора проводилась также по расчету динамических спектрограмм различных звукосочетаний. Однако результаты были не всегда удовлетворительными из-за неточного определения L в динамике речи. После исправления L результаты анализа соответствовали действительности.

Заключение

Описанный выше алгоритм частотно-адаптивного авторегрессионного анализа речевых сигналов хорошо зарекомендовал себя на стационарных речевых фрагментах гласных фонем при правильном рассчитанных значениях длины речевого тракта диктора L . Этот геометрический параметр является очень важным критерием

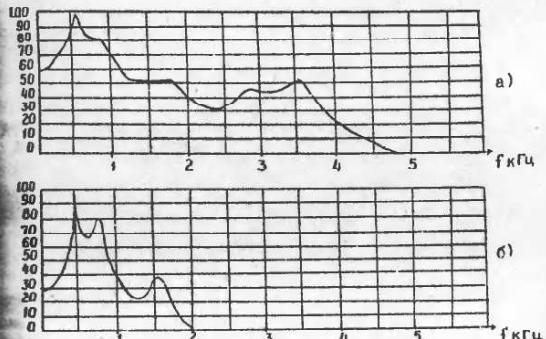


Рис. 7. Спектральные срезы фонемы /O/, полученные:
а) не адаптивным анализатором;
б) частотно-адаптивным анализатором

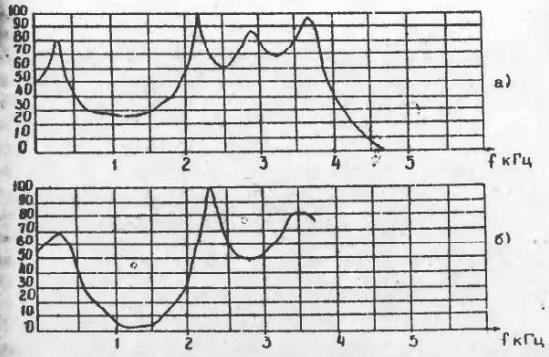


Рис. 8. Спектральные срезы фонемы /И/, полученные:
а) не адаптивным анализатором;
б) частотно-адаптивным анализатором

качеств при адаптации системы к диктору, а так же при изменении геометрии тракта одного и того же диктора во время произношения различных фонем, так как диапазон частот в мгновенном спектре сигнала и длина речеобразующего тракта L коррелируются. Поэтому в структуре анализатора в зависимости от L выбирается тот или иной ФНЧ, рассчитывается соответствующая частота дискретизации и количество отсчетов в кадре, подвергаемом спектральному анализу в данный момент.

ЛИТЕРАТУРА

1. Акинфиев Н.И. К вопросу построения теории речевых сообщений// Сб. научн. труд. ГОС НИИ МРП СССР. - 1957. - Вып. 4. С.3-25.
2. А.с. № 143430 (СССР). Устройство для автоматического слежения за артикуляционными параметрами речи по речевому сигналу с возможностью выделения сигнала-остатка речи/ Акинфиев Н.И.. Опубл. в БИ, 1961. - № 24.
3. Saito S., Itakura F. The Theoretical consideration of Statistically Optimum Methods for Speech Spectral Density Report NJ107. Electrical Communication Laboratory NT.T.-Tokyo, 1966.
4. Atal B.S. Schroeder M.R. Predictive Coding of Speech Signals. Proc. Conf. Commun. and Process., 1967.- Р. 360-361.
5. Маркел Д.Д., Грей А.Х. Линейное предсказание речи. - М.: Связь, 1980.
6. Фан Г. Анализ и синтез речи. - Новосибирск: Наука, 1970.
7. Методы автоматического распознавания речи/ Под ред. Ли У.-М.: Мир, 1983.
8. Paige A., Zue V. Calculation of Vocal Tract Length IEEE Transactions on Audio and Electroacoustics 1970.-Vol.AU-18.
9. Зааль Р. Справочник по расчету фильтров. - М: Радио и связь, 1983.

г. Минск

56

АНАЛИЗ И СИНТЕЗ РЕЧИ

ЧАСТЬ II. СИНТЕЗ РЕЧИ

УДК 621.391

Б.М.Лобанов

МИКРОВОЛНОВОЙ СИНТЕЗ РЕЧИ ПО ТЕКСТУ

Введение

Синтез речи по тексту предполагает наличие определенных процедур (правил) модификации акустических характеристик каждой фонемы в зависимости от ее окружения, позиции в речевой единице, ударения, интонации и других факторов. Поэтому в системах синтеза речи по тексту чаще всего используют формантный синтез сигналов [1,2], позволяющий в широких пределах изменять акустические характеристики звука и таким образом моделировать эффекты коартикуляции, ассимиляции, редукции фонем, управлять мелодическим, ритмическим и динамическим контурами речи. С использованием формантного синтезатора достигается высокое качество синтезированной речи, однако возможности дальнейшего совершенствования ограничиваются в настоящее время неполнотой моделей речеобразования как в целом, так и части моделирования индивидуальных свойств человеческого голоса. В современных формантных синтезаторах практически отсутствует учет взаимодействия источников возбуждения и речевого тракта, динамики изменения формы импульсов возбуждения, индивидуальных свойств речевого тракта и др. Поэтому с помощью формантного синтезатора с трудом удается синтезировать близкий к женскому голос, имитировать эстетически привлекательные голоса или просто копировать любой наперед заданный голос.

Выходом из этого положения могло бы стать использование в системах синтеза речи по тексту отрезков естественной речевой волны. При этом необходимо выбрать в качестве элемента речевой волны такой ее отрезок, который, с одной стороны, позволил бы путем комбинации элементов получать все необходимое многообразие

ние фонетически значимых звуков речи, а с другой - осуществляя их модификацию в соответствии с просодическими правилами. Очевидно, что в качестве этого элемента не могут быть выбраны традиционно-используемые при параметрическом синтезе речи такие элементы, как слоги, дифоны или даже сегменты фонем и аллофонов.

В настоящей работе в качестве элементов естественного речевого сигнала предлагается использовать короткие отрезки волн (микроволн), соизмеримые с периодом основного тона. Сущность микроволнового синтеза заключается в представлении речевого сигнала конечным числом заранее выбранных типов волновых форм (ВФ). Число различных ВФ должно быть выбрано таким, чтобы отразить в значимое разнообразие импульсных реакций вокального тракта в процессе речеобразования. Ориентировочное число ВФ, необходимое для синтеза речи, может быть подсчитано исходя из опыта формального синтеза речевых сигналов. При синтезе звонких звуков используются 3 управляемых по частоте форманты: F_1, F_2, F_3 . Опытным путем установлено, что хорошее качество речи сохраняется при квантовании этих параметров на следующее число градаций: $F_1 = 8, F_2 = 16, F_3 = 4$. Разборчивость синтезированной речи во-еще остается хорошей, если число градаций будет снижено до 4 для F_1 , 8 - для F_2 и 2 - для F_3 . Если допустить, что при синтезе используется все возможное разнообразие значений этих параметров, то получим следующие оценки необходимого числа ВФ:

$$N_{\max} = 8 \times 16 \times 4 = 512;$$

$$N_{\min} = 4 \times 8 \times 2 = 64.$$

Реально при синтезе речи используется не более половины возможных комбинаций, так что требуемое число ВФ для синтеза звонких звуков лежит в пределах от 32 до 256. Для синтеза шумовых звуков необходимо дополнить набор ВФ отрезками шумовых сигналов, число которых лежит в пределах от одного до нескольких десятков.

По-видимому, впервые идея использования в качестве элементов синтеза ВФ, соизмеримых с периодом, высказана в работе [4].

В работе [5] идея синтеза с использованием набора ВФ успешно апробирована в системе дифонного синтеза речи по тексту для мужского и женского голосов. Однако до сих пор не нашли удовлетворительного решения вопросы определения оптимального

набора ВФ, их плавного соединения в текущем речевом потоке, адекватной модификации параметров ВФ в соответствии с просодическими правилами изменения мелодики, ритмики и динамики речи. Решению этих задач, без которых невозможен качественный синтез речи по тексту в полном объеме, посвящена настоящая работа.

I. Микроволновое представление фонем

В основу микроволнового представления фонем положен принцип последовательного разложения фонем на аллофоны и аллофонов - составляющие их сегменты. Возможна различная степень детальности разложения каждой фонемы на аллофоны и аллофонов на сегменты. Здесь мы ограничимся одним из возможных разложений, достаточным как для понимания сущности этой процедуры, так и для обеспечения необходимого многообразия реализаций каждой фонемы аллофоне при синтезе слитной речи.

Из всего множества аллофонов русских гласных фонем (U, O, A, Y, E, I) целесообразно в первую очередь взять их мягкие варианты / U, O, E, I /, а также соответствующие им назализованные варианты / $\bar{U}, \bar{O}, \bar{A}, \bar{E}, \bar{Y}$ и / $\bar{U}, \bar{O}, \bar{A}, \bar{E}, \bar{I}$ /. Каждый аллофон целесообразно представить в виде, по крайней мере, 3 последовательных сегментов: начального, срединного и конечного. При этом тип срединного сегмента (стационарная часть гласной) зависит только от типа выбранного аллофона, а тип начального и конечного сегментов зависит, кроме того, от типа предшествующей и последующей фонем. А также формантного описания начальный, срединный и конечный элементы задают соответственно начало, середину и конец формантных переходов на гласной. Начальный и конечный переходы определяются местом образования предшествующей и последующей фонем [2], так что число типов начального и конечного (переходных) сегментов гласной определяется числом типов фонем, отличающихся местом образования.

Для русских согласных необходимо различать губное, зубное, щекогубное, велярное и латеральное место образования. Таким образом, для описания переходных сегментов каждого аллофона гласной необходимо иметь до 6 различных типов ВФ. Общее представление каждой гласной фонемы в виде набора ВФ, необходимых для писания всех сегментов и аллофонов, дано в табл. I.

Таблица 1

Фонема	/U/				/O/			
	U	Ü	Ǖ	Ǖ̄	U	Ö	Ȫ	Ȫ̄
срединный	W_{11}^u	W_{21}^u	W_{31}^u	W_{41}^u	W_{11}^o	W_{21}^o	W_{31}^o	W_{41}^o
переходн. губной	W_{12}^u	W_{22}^u	W_{32}^u	W_{42}^u	W_{12}^o	W_{22}^o	W_{32}^o	W_{42}^o
переходн. зубной	W_{13}^u	W_{23}^u	W_{33}^u	W_{43}^u	W_{13}^o	W_{23}^o	W_{33}^o	W_{43}^o
переходн. альвеоляр.	W_{14}^u	W_{24}^u	W_{34}^u	W_{44}^u	W_{14}^o	W_{24}^o	W_{34}^o	W_{44}^o
переходн. велярный	W_{15}^u	W_{25}^u	W_{35}^u	W_{45}^u	W_{15}^o	W_{25}^o	W_{35}^o	W_{45}^o
переходн. латеральн.	W_{16}^u	W_{26}^u	W_{36}^u	W_{46}^u	W_{16}^o	W_{26}^o	W_{36}^o	W_{46}^o

Из всего множества аллофонов русских согласных целесообразно в первую очередь взять их варианты, обусловленные эффектом коартикуляции с последующим гласным [2]. При этом для твердых согласных достаточно выделить три аллофона: согласный перед /O/, /U/ - C^u , перед /A/ - C^a и перед /E/, /Y/ - C^e . Кроме того, выделяется один вариант мягких согласных C' . Для каждого аллофона согласных определяются три временных сегмента - начальный, срединный и конечный. Таким образом, описание каждого согласного в виде набора ВФ может быть представлено с помощью табл. 2.

Описанное представление русских фонем в виде ВФ сегментов аллофонов допускает значительное изменение их количества как в сторону увеличения, так и в сторону уменьшения в зависимости от предъявляемых требований к качеству синтезированной речи либо к объему запоминаемой информации.

Таблица 2

Фонема	/Z/				/L/			
	Z ^e	Z ⁱ	Z ^a	Z ^u	L ^e	L ⁱ	L ^a	L ^u
срединный	W_{11}^z	W_{21}^z	W_{31}^z	W_{41}^z	W_{11}^l	W_{21}^l	W_{31}^l	W_{41}^l
начальный	W_{12}^z	W_{22}^z	W_{32}^z	W_{42}^z	W_{12}^l	W_{22}^l	W_{32}^l	W_{42}^l
конечный	W_{13}^z	W_{23}^z	W_{33}^z	W_{43}^z	W_{13}^l	W_{23}^l	W_{33}^l	W_{43}^l

2. Соединение ВФ в речевом потоке

Для синтеза звонких звуков используются ВФ, вырезанные из речевого сигнала на соответствующих стационарных и переходных участках звуков, с длительностью, равной периоду основного тона. Соединение ВФ на стационарных участках сводится просто к их последовательному считыванию. Для переходных участков такая процедура могла бы подойти только в случае предварительной записи нескольких ВФ на каждый тип переходного участка. Это практически невозможно осуществить как в плане трудоемкости подготовки и прарирования исходного речевого материала, так и в плане чрезмерного разрастания требуемого объема памяти и количества правил синтеза переходных участков.

Существует одна интересная возможность обойти указанные трудности, основанная на использовании инерционных свойств слухового восприятия человека, аналогичных зрительному. Хорошо известно, что впечатление плавного замещения одной слайд-картишки другой может быть достигнуто путем плавного уменьшения яркости (от исходной до нуля) одного изображения и одновременного увеличения яркости (от нуля до необходимой) другого изображения, спроектированных на один и тот же экран. Проведенные нами исследования показали, что аналогичный эффект замещения присущ и слуховому восприятию звуков. Слуховой эффект плавного замещения достигается путем создания интервала перекрытия двух звуков с постепенным уменьшением амплитуды первого звука и одновременным увеличением амплитуды второго на интервале перекрытия (рис. I, а).

В результате суммирования (рис. I, б) в звуковом поле на участке перекрытия образуется сложный звук, воспринимаемый как плавный переход от первого ко второму звуку. В какой-то мере плавность этого перехода фиксируется и на сонограмме (рис. I, в).

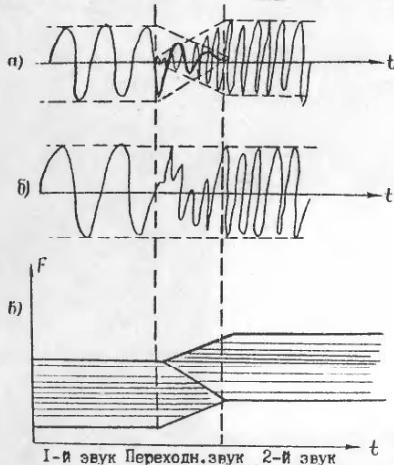


Рис. I

Математическое описание алгоритма, реализующего механизм плавного замещения последовательности ВФ, может быть дано следующим образом. Пусть для определенности требуется осуществить плавный переход на интервале гласной фонемы от начальной волны W_{12} к средней W_{11} и затем к конечной W_{13} . Это соответствует (см.табл. I) переходному процессу из гласной в сочетании: "губная согласная - гласная - зубная согласная" (например, в слове "PAT").

Обозначим длительности начального перехода T_{12} , срединного стационара T_{11} и конечного перехода T_{13} . Требуемые значения длительностей берутся из таблиц, подобных табл. I, 2. В соответствии с изложенным выше правилами плавного замещения ВФ на участке гласной необходимо сформировать два сигнала:

$$S_1(t) = \begin{cases} W_{12}(t) \frac{1}{T_{12}} t, & 0 \leq t \leq T_{12} \\ W_{11}(t), & T_{12} < t < T_{12} + T_{11} \\ W_{11}(t) \frac{1}{T_{12}} (T_{12} - t), & T_{12} + T_{11} < t < T_{12} + T_{11} + T_{13}, \end{cases} \quad (1)$$

$$S_2(t) = \begin{cases} W_{12}(t) \frac{1}{T_{12}} (T_{12} - t), & 0 \leq t \leq T_{12} \\ 0, & T_{12} < t < T_{12} + T_{11} \\ W_{13}(t) \frac{1}{T_{13}} (t - T_{12} - T_{11}), & T_{12} + T_{11} \leq t \leq T_{12} + T_{11} + T_{13}. \end{cases} \quad (2)$$

Огибающие этих сигналов изображены на рис. 2.

Сигнал $S(t)$, в котором реализуется эффект плавного замещения последовательности ВФ W_{12}, W_{11}, W_{13} на интервале звучания гласной фонемы, образуется из сигналов (1), (2) путем простого суммирования: $S(t) = S_1(t) + S_2(t)$.

Исследования, проведенные с использованием различных типов ВФ, показали, что слуховой эффект их плавного замещения возникает лишь в том случае, когда соответствующий им переходный процесс в естественной речи характеризуется одновременным линейным движением формантных частот (рис. 3, а). Если же в естественной речи при переходе от ВФ W_1 к W_2 (рис. 3, б) движение формантных частот не удовлетворяет этому условию, то для достижения слухового эффекта плавного замещения ВФ $- W_1$ и W_2 необходимо ввести промежуточную ВФ $- W_{12}$. В частности, рис. 3, а иллюстрирует переход от гласной /A/ к /E/, а рис. 3, б - от /A/ к /I/.

3. Управление просодическими параметрами

Просодика речи (мелодика, динамика, ритмика) задается путем текущего управления частотой основного тона, амплитудой и длительностью звуков. Рассмотрим особенности и способы управления этими параметрами при микроволновом способе синтеза.

Простейшим способом управления частотой основного тона является следующий. Пусть исходная ВФ имеет длительность T'_0 , причем T'_0 выбрана внутри определяемого просодическими правилами диапазона изменения длительности периодов основного тона:

$$T_{\text{мин}} < T'_0 < T_{\text{макс}}$$

В качестве конкретного значения T'_0 может быть взято среднестатистическое значение периода основного тона речи диктора, используемого при формировании набора ВФ. Если текущее значение $T_0 = T'_0$,

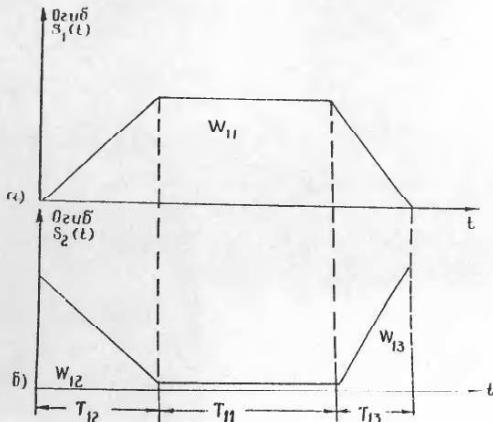


Рис. 2

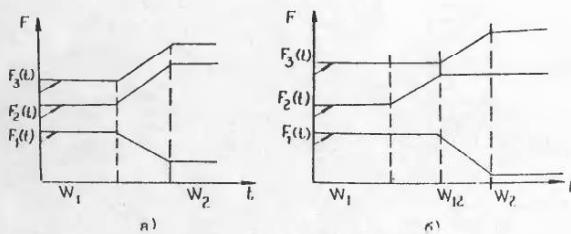


Рис. 3

то речевой сигнал образуется путем простого повторения заданий ВФ (рис.4,а). При $T_0 > T'_0$ повторное считывание ВФ начинается спустя временной интервал $T_0 - T'_0$, в сам интервал заполняется нулями. При $T_0 < T'_0$ в момент $t = T_0$ мгновенно прекращается считывание ВФ и начинается процесс ее повторного считывания.

Экспериментальное исследование данного метода управления частотой основного тона показало, что он обеспечивает достаточно высокое качество синтезированного звука. Это определенно можно утверждать для случая $T_0 > T'_0$, если интервал $T_0 - T'_0$ не превышает 30% от длительности периода T_0 . Для случая $T_0 < T'_0$ искажения не заметны на слух, если момент прекращения считывания приходит на значение ВФ близи нуля (10% - 20% от амплитуды ВФ). В противном случае (см.рис.4,б) наблюдается отчетливое искажение звучания, напоминающее назализацию (гнусавость). Этот дефект может быть устранен путем соответствующего слаживания процесса резкого прекращения считывания. Это может быть достигнуто, например, следующими двумя способами. В первом способе в моменты начала повторного считывания включается фильтр 2-го порядка с постоянной времени $\tau = 1/4 T_0$ (рис.5,б). Во втором способе перед началом повторного считывания ($\tau = 1/4 T_0$) сигнал ВФ умножается на гладкую единичную функцию (рис.5,в). Это может быть, например, функция вида

$$y = e^{-\alpha t^2}$$

Использование одного из этих методов для случая $T_0 < T'_0$ дает вполне удовлетворительные результаты. Для случая $T_0 > T'_0$ можно использовать способ дополнения периода нулями (см.рис.4,б) при условии, что выбирается $T'_0 = 0,7 T_{0 \max}$.

Рассмотрим далее особенности управления двумя различными просодическими характеристиками: длительностью и амплитудой звуков. Определяемая ритмическим контуром требуемая длительность звонких звуков может быть получена путем задания необходимого числа периодов на каждом фонемном сегменте. Если требуемая длительность i -го фонемного сегмента равна T_i , то в среднем для ее реализации необходимое число \bar{n}_i периодов определяется по формуле

$$\bar{n}_i = \frac{T_i}{T_{vi}} ,$$

где T_{oi} - среднее значение длительности периода основного тона на i -м фонемном сегменте, задаваемого мелодическим контуром. Точное значение необходимого числа периодов определяется путем сравнения требуемой длительности сегмента T_i и суммарной текущей длительности периодов основного тона, реализуемых на этом сегменте. При этом длительность сегмента задается с погрешностью, равной длительности последнего периода основного тона. Длительность шумных звуков задается путем простого подсчета необходимого количества отсчетов соответствующего шумового сигнала.

Последний просодический параметр - текущая амплитуда звука - задается путем умножения последовательности отсчетов звука на текущие значения динамического контура.

4. Синтез последовательности ВФ по фонемному тексту

Для синтеза речи по фонемному тексту необходимо реализовать процедуру генерации последовательности ВФ, описывавших каждую фонему с учетом ее текстового окружения. Процедуру генерации ВФ удобно описать, базируясь на понятии портрета фонемы [2]. Под портретом фонемы понимается некоторое универсальное описание, заданное в виде набора констант или некоторых функций, достаточное для генерации множества ее временных сегментов с учетом позиционно-комбинаторной изменчивости. В отличие от формантных портретов фонем [2] волновые портреты могут быть описаны существенно более компактно. Это связано с тем, что в ВФ уже заложена информация о формантных характеристиках звука. В волновом портрете фонемы нужно лишь указать, какой конкретной ВФ W^j описывается тот или иной сегмент фонемы, на какой длительности T и какой амплитуде A^j необходимо задать выбранную ВФ для описания данного сегмента, а также за какое время τ^j должно устремиться стационарное значение ВФ для данного сегмента. Обобщенный волновой портрет фонемы, задаваемый на трех последовательных временных сегментах, представлен в табл.3.

Из табл.3 видно, что конкретный выбор ВФ на каждом сегменте определяется не только типом текущей синтезируемой фонемы, но и ее непосредственным окружением в тексте. В общем случае выбор требуемой ВФ осуществляется с помощью многозначной функции

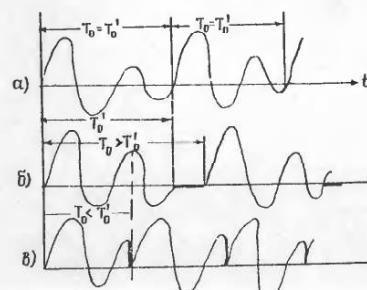


Рис. 4

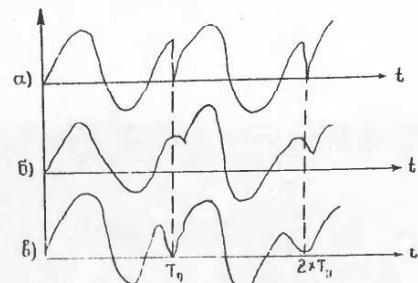


Рис. 5

$$W_i^j = \begin{cases} W_{11}^j & \text{при } p_{i-1} \in \Phi_1, p_{i+1} \in \Phi_1; \\ W_{12}^j & \text{при } p_{i-1} \in \Phi_1, p_{i+1} \in \Phi_2; \\ W_{21}^j & \text{при } p_{i-1} \in \Phi_2, p_{i+1} \in \Phi_1; \\ \dots & \dots \\ W_{m,n}^j & \text{при } p_{i-1} \in \Phi_m, p_{i+1} \in \Phi_n; \\ \dots & \dots \end{cases} \quad (3)$$

Здесь W_i^j - ВФ, необходимая для синтеза j -го сегмента i -й фонемы текста; p_{i-1} , p_{i+1} - предшествующая и последующая фонемы текста; $\Phi_m\Phi_n$ - множества фонем m -го и n -го типов.

Формула (3) приобретает конкретный вид для каждой фонемы и ее сегментов, если учесть правила микроволнового представления фонем, изложенные в разд. I.

Таблица 3

Сегмент Параметр	начальный	срединный	конечный
W_i^1	$f_w^1(p_i \pm 1)$	$f_w^2(p_i \pm 1)$	$f_w^3(p_i \pm 1)$
T_i^1	$f_T^1(p_i \pm 1)$	$f_T^2(p_i \pm 1)$	$f_T^3(p_i \pm 1)$
A_i^1	$f_A^1(p_i \pm 1)$	$f_A^2(p_i \pm 1)$	$f_A^3(p_i \pm 1)$
V_i^1	$f_V^1(p_i \pm 1)$	$f_V^2(p_i \pm 1)$	$f_V^3(p_i \pm 1)$

Приведем для примера конкретный вид формулы (3) для трех сегментов гласной фонемы /A/. Для начального сегмента имеем

$$W_i^1 = \begin{cases} W_1^1 & \text{при } p_{i-1} \in \Phi_1 \\ W_2^1 & \text{при } p_{i-1} \in \Phi_2 \\ W_3^1 & \text{при } p_{i-1} \in \Phi_3 \\ W_4^1 & \text{при } p_{i-1} \in \Phi_4 \\ W_5^1 & \text{при } p_{i-1} \in \Phi_5 \\ W_6^1 & \text{при } p_{i-1} \in \Phi_6 \\ W_7^1 & \text{при } p_{i-1} \in \Phi_7 \\ W_8^1 & \text{при } p_{i-1} \in \Phi_8 \\ W_9^1 & \text{при } p_{i-1} \in \Phi_9 \\ W_{10}^1 & \text{при } p_{i-1} \in \Phi_{10} \\ W_{11}^1 & \text{при } p_{i-1} \in \Phi_{11} \end{cases} \quad (4)$$

Здесь $\Phi_1 = \{P, B, F, V, L\}$ - множество твердых губных и боковых согласных, $\Phi_2 = \{T, D, S, Z, R, C, CH, ZH, K, G, X\}$ - множество зубных, альвеолярных и небных твердых согласных, $\Phi_3 = \{P', B', F', V'\}$ - множество губных мягких согласных, $\Phi_4 = \{T', D', S', Z', R', CH', ZH'\}$ - множество мягких зубных и альвеолярных согласных, $\Phi_5 = \{K', G', X'\}$ - множество мягких небных согласных, $\Phi_6 = \{L'\}$ - единичное множество мягких боковых согласных, $\Phi_7 = \{M\}$ - единичное множество твердых губных носовых согласных, $\Phi_8 = \{N\}$ - единичное множество твердых зубных носовых согласных, $\Phi_9 = \{M'\}$ - единичное множество мягких зубных носовых согласных, $\Phi_{10} = \{N'\}$ - единичное множество мягких губных носовых согласных, $\Phi_{11} = \{U, O, A, E, Y, \#\}$ - множество гласных и паузы.

Для срединного сегмента /A/ имеем

$$W_i^2 = \begin{cases} W_1^2 & \text{при } p_{i-1} \in \Phi_{12} \\ W_2^2 & \text{при } p_{i-1} \in \Phi_{13} \\ W_3^2 & \text{при } p_{i-1} \in \Phi_{14} \\ W_4^2 & \text{при } p_{i-1} \in \Phi_{15} \end{cases} \quad (5)$$

Здесь $\Phi_{12} = \{Ц, Ъ, А, Е, У, Л, Р, В, З, Ч, Н, К, С, Ф, К, С, Ш, Х, С\}$ - множество твердых неносовых согласных и гласных, $\Phi_{13} = \{J, Л', Р', В', З', Б', Д', Г', Р', Т', К', Ф', С', Ш', Х', Ч\}$ - множество мягких согласных, $\Phi_{14} = \{М', Н'\}$ - множество мягких носовых согласных, $\Phi = \{М, Н\}$ - множество твердых носовых согласных.

Для конечного сегмента имеем

$$W_i^3 = \begin{cases} W_1^3 & \text{при } p_{i+1} \in \Phi_1 \\ W_2^3 & \text{при } p_{i+1} \in \Phi_2 \\ W_3^3 & \text{при } p_{i+1} \in \Phi_3 \\ W_4^3 & \text{при } p_{i+1} \in \Phi_4 \\ W_5^3 & \text{при } p_{i+1} \in \Phi_5 \\ W_6^3 & \text{при } p_{i+1} \in \Phi_6 \\ W_7^3 & \text{при } p_{i+1} \in \Phi_7 \\ W_8^3 & \text{при } p_{i+1} \in \Phi_8 \\ W_9^3 & \text{при } p_{i+1} \in \Phi_9 \\ W_{10}^3 & \text{при } p_{i+1} \in \Phi_{10} \\ W_{11}^3 & \text{при } p_{i+1} \in \Phi_{11} \\ W_{12}^3 & \text{при } p_{i+1} \in \Phi_{12} \\ W_{13}^3 & \text{при } p_{i+1} \in \Phi_{13} \\ W_{14}^3 & \text{при } p_{i+1} \in \Phi_{14} \\ W_{15}^3 & \text{при } p_{i+1} \in \Phi_{15} \end{cases} \quad (6)$$

Здесь множества $\Phi_1 - \Phi_{10}$ те же, что в формуле (4), а множества $\Phi_{11}, \Phi_{12}, \Phi_{18}, \Phi_{19}, \Phi_{20}$ являются единичными и содержат соответственно гласные /и. ы. а. е. у/.

С учетом конкретных правил аллофонической изменчивости аналогичные формулы выбора ВФ сегментов могут быть записаны для каждой фонемы.

Выбор длительности и амплитуды сегментов фонем осуществляется по тем же формулам, что и выбор ВФ, так что каждой ВФ может быть сопоставлена длительность ее повторения и амплитуда.

Как уже указывалось, процесс установления ВФ на каждом сегменте может быть реализован путем моделирования слухового эффекта плавного замещения ВФ. Выбор времени замещения t^j осуществляется по формулам, аналогичным (но не идентичным) формулам для выбора ВФ. Для гласных фонем $t_f^1 T, t_f^2 0, t_f^3 = T^3$, т.е. первый сегмент замещается вторым за время, равное длительности первого сегмента, на втором сегменте процесс замещения отсутствует, а на интервалах длительности третьего сегмента осуществляется процесс замещения второго сегмента третьим (рис.6). Для согласных фонем устанавливаются фиксированные значения t_c^1, t_c^2, t_c^3 , зависящие только от типа согласной фонемы. При этом t_c^1 устанавливает время замещения ВФ последнего сегмента ($i+1$ -й) фонемы первым сегментом i -й согласной фонемы, t_c^2 устанавливает время замещения первого сегмента вторым и третьим. Особая роль отведена t_c^3 , которое определяет опережающее замещение третьего сегмента $i+1$ -й согласной фонемы первым сегментом ($i+1$ -й гласной фонемы) и реализуется только в случае, если ($i+1$)-я фонема является гласной (см.рис.6).

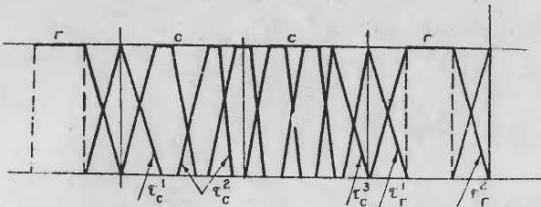


Рис. 6

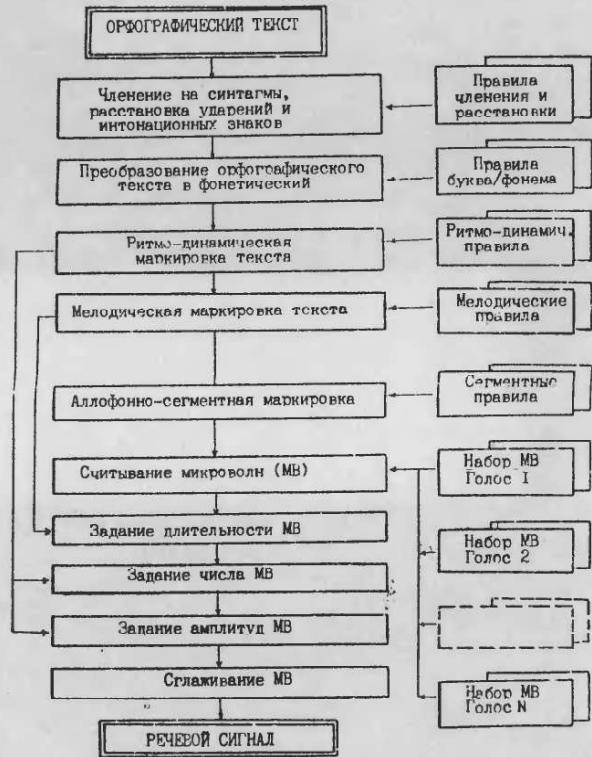


Рис. 7. Блок-схема метода микроволнового синтеза речи

Заключение

Алгоритм микроволнового синтеза речи

Блок-схема алгоритма приведена на рис.7. Орфографический текст, поступающий на вход системы микроволнового синтеза речи, расчленяется на синтаксы и размечается ударениями и знаками интонации. Затем он преобразуется в фонетический текст, который, в свою очередь, размечается в соответствии с ритмо-динамическими и фонетическими правилами. Совокупность этих правил позволяет в конечном итоге задать необходимые длительности, амплитуды и период основного тона синтезируемых фонем. Фонетический текст в соответствии с сегментными правилами преобразуется в последовательность аллофонов, а затем – сегментов, которые определяют результатирующую последовательность микроволн, соответствующую исходному орфографическому тексту. На участках переходов между фонемами производится необходимое сглаживание соседних микроволн. Сформированная таким образом последовательность микроволн через ЦПУ выводится на динамик.

Система позволяет хранить микроволны для различных голосов (мужских и женских), поэтому синтез необходимого голоса можно осуществить путем записи и считывания соответствующего набора микроволн. Заметим, что при этом в значительной степени сохраняются индивидуальные особенности, присущие голосу каждого диктора.

Идея микроволнового синтеза речи была в полном объеме проявлена на примере синтеза русской речи. Сейчас ведутся работы по синтезу белорусской, словацкой, чешской, английской, немецкой и монгольской речи. Основной проблемой при микроволновом синтезе речи на новом языке является проблема адекватного выбора инвентаря аллофонов и составляющих их сегментов. На этом этапе совершенно необходимо участие квалифицированного эксперта-фонетиста, а также носителя нормативного языка с хорошей дикцией и приятным голосом. Блок-схема алгоритма многоголосичного синтеза остается неизменной, изменяются только наборы правил и наборы необходимых микроволн.

ЛИТЕРАТУРА

1. Klat D. The klettalk text-to-speech conversation system. Proc. IEEE ICASSP, Paris, 1982.
2. Lobanov B.M. The Phonemafons text-to-speech system. Proc. ICPHS, Tallinn, 1987.
3. Morel M. Synthese vocale par recordement de segment d'oscillogrammes Revue d'Acoustique, vol.14, no 56, 1981.
4. Hamon G. Synthèse par concaténation de formes d'ondes. Note technique, NT/LIA/TSS/355 CNET, 1982.

г. Минск

АНАЛИЗ И СИНТЕЗ РЕЧИ

УДК 621.320

Г.В.Лосик

КОМПЬЮТЕР С СИНТЕЗОТОРОМ РЕЧИ ДЛЯ СЛЕПЫХ

Проблема

В настоящее время в мировой практике широкое распространение получили компьютеры для социально-трудовой реабилитации инвалидов по зрению, рационального труда и трудоустройства работников интеллектуального труда, обучения незрячих студентов и школьников. Первым шагом на этом пути было создание в 70-е гг. калькуляторов с брайлевским выходом (Япония, США). В СССР их опытные образцы были разработаны в 80-е гг. в Свердловске. Затем наряду с брайлевским выходом в калькуляторах появился вывод информации с помощью устной речи. Такой аппарат, как "ВерсоБрайль" (ЭРГ) реализует кроме калькулятора функции пишущей брайлевской машинки с памятью.

С появлением персональных ЭВМ у незрячих людей значительно расширились возможности использования своих интеллектуальных способностей. Компьютерное оборудование существенно уменьшает отличие в работе с ПЭВМ незрячего и зрячего пользователя. Необходимость специального тифлотехническом оборудовании почти отпадает. Из несерийного оборудования необходимы только брайлевские принтер линейка-дисплей.

В СССР примерно 200000 инвалидов по зрению: totally слепых и слабовидящих. В настоящее время участь 70% из них - заниматься рутинной манипуляционной ручной работой. В то же время около 60000 инвалидов имеют среднее специальное и высшее образование. Вторую группу численностью около 30000 человек составляют учащиеся средних и специальных школ. Анализ показывает, что большинство слепых сегодня потенциально готово

74

к работе с персональным компьютером. Реабилитация этих людей как специалистов интеллектуального труда возможна именно за счет уравнивания их со зрячими в доступе к компьютерным базам данных, машинным банкам, фондам алгоритмов и программ. Вместе с тем ввиду оснащения сегодня многих вузовских курсов персональными ЭВМ ряд институтов и отдельные факультеты страны прекратили прием слепых абитуриентов. В итоге некоторые современные специальности для слепых людей скоро могут стать недоступными.

Как и зрячие, слепые учащиеся старших классов в специшко-ле изучают информатику. Например, в Белоруссии в обычных школах сегодня уже около 1000 компьютеров. Однако слепые учащиеся в настоящее время в школе лишены возможности работать даже с обычным калькулятором. Известно, что некоторые слепые становятся программистами, овладевают языками программирования, широко пользуются ЭВМ. К этому их обязывает научный прогресс. Однако они работают на ЭВМ в "пакетном режиме" или со зрячим оператором, т.е. не в диалоге с дисплеем и клавиатурой. Такие незрячие программисты работают в Москве, Ленинграде, Новосибирске. Их пример служит своеобразным экспериментом, подтверждющим, что врожденная или приобретенная патология зрения не оказывается на способности человека овладевать языками программирования. Более того, опыт показывает, что слепой человек, будучи ущемленным в способности овладевать письменной формой человеческого языка, оказывается наделенным по закону психолого-ической компенсации гипервысокой способностью к овладению машинными языками, пишет программы с необычайно малым числом ошибок.

Существует мнение, что в компьютере слепому недоступна клавиатура. Но практика показала, что на клавиатуре перфоратора или пишущей машинки многие могут обучиться работать вслепую. И поэтому брайлевские точки на клавишах становятся не обязательными. Следовательно, единственным препятствием в работе является неспособность слепого читать экран дисплея. Поэтому именно синтезатор речи дает выход из кризисного положения. Он позволяет синтезировать ис строкам текст, который компьютер выдает на экран. В практике у слепых уже давно существует

75

способ работы "с чтецом", когда зрячий читает с листа слепому ту или иную статью журнала, книгу. Последний слушает, время от времени останавливает чтеца, перестраивает его. Синтезатор речи, подключенный к дисплею, становится своеобразным артистом, заменяющим чтеца. Если быть более точным, синтезатор речи заменяет не только чтеца, но и дисплей, экран которого для незрячего становится излишним. В составе компьютера может оставаться процессор с дисководами и клавиатурой. Поэтому в размерах и стоимости такая ЭВМ сокращается почти вдвое.

Появление синтезатора устной речи совершило революцию в компьютеризации незрячих. Вывод из компьютера информации через брайлевскую строку оказался эргономически менее удобным, чем вывод через синтезатор речи. Устноречевое восприятие

- a) более быстрое, чем тактильное;
- b) более информативное в связи с наличием просодии;
- c) не требует специального дополнительного навыка распознавать шестиглавчики кончиками пальцев;
- d) освобождает пальцы слепого для клавиатуры и позволяет оператору выполнять сенсорные и моторные действия параллельно.

Разработки в области сервисного оснащения ПЭВМ синтезаторами речи начались в 70-е гг. В ряде стран наложен серийный выпуск ПЭВМ с синтезаторами речи. Однако в таких компьютерах роль синтезатора остается факультативной, в дисплее — обязанностью, а не наоборот. Поэтому для слепых людей они малоприемлемы. В США, Японии, Италии, ФРГ наложен выпуск компьютеров и калькуляторов для слепых. Их отличает присутствие специального программного обеспечения для "слепой" диалоговой работы оператора с ПЭВМ, словарей для расстановки ударений, наличие баз данных. Однако нерусскоговорящие синтезаторы речи и связанные с ними программное обеспечение нельзя заменить и сделать русскоговорящими.

В Советском Союзе пока нет серийно выпускаемых калькуляторов или компьютеров для слепых. Проведенный нами поисковый анализ показал, что на отечественных и зарубежных выставках демонстрируются калькуляторы и спецкомпьютеры для слепых, разработанные в Японии, США, Италии. Наряду с "брайлевским" выводом информации в них реализуется устноречевой вывод с помощью

синтезированной речи. Один из компьютеров для слепых, разработанный в США, демонстрировался на выставке "Информатика в жизни США" в 1989 г. Компьютер снабжен синтезатором английской речи и "брайлевским" принтером на плотную бумагу. В НИИ медицинского приборостроения в отделе тифлотехники разработан брайлевский спектрограф, рассчитанный на присоединение к любой серийной ПЭВМ. Это линейка брайлевских шестиглавочных модулей. Она позволяет слепому работать с ПЭВМ, воспринимая текст построчно кончиками пальцев по Брайлю.

Решение

В лаборатории анализа и синтеза речи Института технической кибернетики АН БССР проведена разработка персонального компьютера для русскоговорящих слепых. Разработка выполнена на базе серийных ПЭВМ типа EC 1840-41, IBM PC и синтезатора речи Фонемофон-5. Последний может реализовать 5 диакторских голосов, 300% изменения темпа речи, 400% громкости. С его помощью возможно чтение стихов и исполнение песен. Фонемофон представляет собой плату внутри ПЭВМ. Обращение к синтезатору осуществляется на языке СИ или Ассемблере, с клавиатурой или файла. Синтезатор реагирует на знаки ! ? , . пробел, на цифры от 0 до 9000000, на буквы от А и а до Я и я. Лингвистический процессор синтезатора самостоятельно с письменного текста формирует устный, исправляет около 20 типов ошибок, в 98% случаев автоматически ставит ударение. Для подготовки к аудированию годится любой компьютерный редактор текста.

В математическом плане разработка указанного компьютера осуществлена путем создания специальных пакетов программ, эргономического изучения "слепого" диалога, повышения надежности восприятия синтезированной речи.

Программное обеспечение компьютера для слепых

Известный нам зарубежный опыт в большей или меньшей степени применен при разработке программы для русскоговорящих компьютеров в зависимости от разных социальных и возрастных категорий незрячих пользователей: учащихся специал., программистов, редакторов текстов, работников информационных служб, пользов-

вателе' игровых программ.

В тех случаях, когда на синтезатор осуществляется вывод больших русских текстов ("звучатая книга"), зарубежные синтезаторы в компьютерах для слепых оказываются малоэффективными. Синтез русского слова в них осуществляется "склейкой" звучания слов и букв, но не по формантам, как это сделано в Фонемофоне. Поэтому моделирование интонации вопроса, повествования, восклицания, перечисления в них невозможно. Здесь отсутствует задание словесного и фразового ударения. Однако при работе на таких синтезаторах в режиме отладки программ на Бейсике, СИ, Фортране синтез небольшого лексикона русских слов может достигаться с удовлетворительным качеством.

В зарубежных компьютерах для слепых в большинстве случаев синтезатор речи реализован программно на винчестере. Хотя он занимает в ОЗУ много места, однако транзактировать его легко. В то же время программная громоздкость такого синтезатора предрекает путь к внедрению его на отечественных ПЭВМ, в которых нет винчестера, а ОЗУ меньше 1 Мб., и вынуждает приобретать для него дефицитные ПЭВМ серии IBM PC или EC 1842.

Проведенный нами анализ показал, что для учащихся специальных школ необходимы программы "Обучение клавиатуре", "Адаптация к синтезированной речи", "Калькулятор", всевозможные программы по биологии, русскому языку, информатике. На наш взгляд, компьютер в школе слепых и слабовидящих следует внедрять прежде всего в качестве заменителя чтеца для самообучения школьника как младших, так и старших классов широкому кругу дисциплин. Обучение же информатике и программированию в старших классах необходимо рассматривать как частный случай использования ПЭВМ.

Для незрячих студентов вузов и техникумов, широкого круга людей, планирующих иметь компьютер в личном пользовании, необходимы, кроме того, ППП "Записная книжка", "Справочник", "Календарь", "Пишущая машинка", "Библиотека игр". В вышеперечисленных случаях синтезатор компьютера работает главным образом в режиме синтеза больших текстов. Здесь существенна реализация разных видов интонации, пауз, темпа, необходима быстрая пропстановка ударения в словах озвучиваемого текста. В то же время устноречевое обслуживание вычислительного режима работы ком-

пьютера здесь не обязательно.

Для третьей категории пользователей - работников интеллектуального труда, профессиональных программистов - необходимы ППП "Бейсик", "Паскаль", "СИ", энциклопедические справочники. Для таких пользователей необходима, во-первых, мощная ПЭВМ с винчестером и большим по объему ОЗУ и, во-вторых, специально переделанная на устноречевой лексикон и адаптированная для синтезатора речи операционная система МС ДОС. Такая переделка операционной системы совершена в некоторых зарубежных компьютерах для слепых. Высокое качество синтезированной речи для озвучивания лексикона МС ДОС в 200 - 300 слов не требуется. Поэтому такие русскоязычные компьютеры, оснащенные брайлевским принтером, могут удовлетворять запросам программистов при условии отсутствия в их работе необходимости в озвучивании больших текстов.

В настоящее время в Институте технической кибернетики АН БССР создан АРМ для слепого с минимальной конфигурацией ППП: "Обучение клавиатуре", "Калькулятор", "Записная книжка", "Пишущая машинка", "Словари", "Часы". Синтезатор речи обеспечивает "слепой" диалог с IBM PC AT через обычную клавиатуру. Вопрос, ответ, предупреждение, ошибка, итог работы маркируются 5 различными дикторскими голосами, сменой темпа и громкости речи.

При работе с программой "Обучение клавиатуре" у оператора формируется двигательный навык быстрого нахождения вслепую и нажатия нужной клавиши на клавиатуре, навык быстрого слухового контроля через синтезатор правильности нажатия клавиши, навык быстрого исправления ошибки.

Программа "Калькулятор" позволяет незрячему оперативно выполнять на компьютере всевозможные математические операции с числами при выводе результатов в устной форме через синтезатор.

Программа "Записная книжка" обеспечит оперативное воспроизведение в устной форме созданных на дискетах текстовых файлов-записей, позволит слепому самому вводить с клавиатуры текстовые записи.

При загрузке программы "Пишущая машинка" слепой человек может использовать компьютер как пишущую машинку с русским или латинским шрифтом, с заглавными и прописными буквами, чи-

рами, с запоминанием напечатанных текстов, их редактированием и распечаткой на бумагу, составлением томов, архивов, описей.

В пакете прикладных программ имеется, кроме того, "Англо-русский словарь" на 5000 слов, расписание поездов, часов.

Особой актуальностью отличается программа "Звучащая книга". Ее разработка лишь начата и требует отдельной сложной подпрограммы автоматической простановки слогесного ударения. Программа "Звучащая книга" сможет заменить слепому его чтение, обеспечивая озвучивание через синтезатор любых текстов на русском и английском языках, машинные фонды которых в настолющее время в нашей стране уже огромны.

С помощью перечисленного базового набора программы можно создать АРМы для работы слепых по специальностям юриста, будущего галтера, делопроизводителя, работника архива, информационной службы, научного работника, диспетчера, учетчика, программиста, машинистки, библиотекаря, секретаря-машинистки. Если расширить базовый набор программ, то копьютер для слепых сможет применяться в школе для обучения незрячих. При этом синтезатор речи Фонемофон может работать на базе ПЭВМ Карвет, ДВК-3, Ямха, Нейрон и т.д.

Г. Минск

АНАЛИЗ И СИНТЕЗ РЕЧИ

УДК 621.352

С.С.Сихеев

ИНЖЕНЕРНО-ПСИХОЛОГИЧЕСКИЕ ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ СИНТЕЗАТОРА РЕЧИ НА СПРАВОЧНОЙ СЛУЖБЕ "09"

В настолющее время в нашей стране и за рубежом наложен выпуск нескольких типов синтезаторов устной речи. Наиболее перспективным является реализуемый в них метод синтеза по печатному тексту. Он позволяет озвучивать "машинной" речью любые цифровые и буквенные тексты, которые записаны в ЭВМ. Компьютер, снаженный таким синтезатором, становится своеобразным речевым автоматом, заменяющим диктора, чтеца, телефонистку. Благодаря синтезатору любая текстовая или цифровая информация (произвольный файл, слова, фразы, абзацы, таблицы), которая выводится на экран дисплея, превращается в устную речь и может быть подана человеку через громкоговоритель, наушники или на расстояние по телефону.

В итоге синтезатор речи с позиций инженерной психологии превращается в новое перспективное средство отображения информации (СОИ), перераспределяющее нагрузку со зрительного анализатора человека на слухоречевой. В связи с появлением этой технической новинки возникает проблема инженерно-психологической грамотности ее использования в конкретных диалоговых системах "человек-ЭВМ". При этом становится важным выбор "древовидной" структуры клавиатурно-устикоречевого диалога человека и синтезатора, а также выбор темпа представления синтезированной речи и ее лексикона.

В синтезаторе речи "Фонемофон -WS", разработка которого осуществлена в Институте технической кибернетики АН БССР, предложены 3 варианта "машинного" голоса для двух языков - рус-

Таблица

# п/п	Типовая ситуация	Стандартизованный ответ телефонистки
1	Сложная справка	Переключаю Вас на дежурную. Пожалуйста, подождите
2	Запрос абонента не входит в компетенцию службы "09"	Извините, служба "09" выдает ТОЛЬКО НОМЕРА ТЕЛЕФОНОВ. За справками об адресах, часах работы организаций и другими сведениями обращайтесь в киоск горсправки
3	По запросу абонента в базе данных нет запрашиваемого номера телефона	Простите, по Вашим данным на службе "09" нет информации
4	Запрос абонента не точен и не содержит полной информации	Простите, для удовлетворения Вашего запроса необходима полная информация. Пожалуйста, уточните исходные данные
5	Абонент не может дозвониться кому-либо из-за невнимательного набора номера телефона или ошибочного соединения	Простите, названный Вами номер не менялся. Пожалуйста, дозванивайтесь

Использование синтезатора речи позволяет не только сократить непроизводительную загрузку телефонных каналов (см. ситуации № 2 и № 4), но и снизить нервно-психическую напряженность телефонисток, которую создают, например, диалоги в ситуации № 5.

По дисплейным рабочим местам и клавиатуре имеются следующие замечания:

I. Наличие латинского юникода, не используемого в работе.

ского и английского, 5 вариантов темпа речи, имеется интонация вопроса, восклицания и повествования, предусмотрен режим повтора. Такие технические возможности данного синтезатора речи позволили нам путем соответствующей инженерно-психологической его доработки создать проект внедрения данного синтезатора на городской телефонной сети.

В связи с ростом телефонизации городов, а также частым изменением номеров телефонов организаций и частных лиц нагрузка на справочные службы города в последнее время возросла. Приблизительно 40% абонентов не могут дозвониться до службы "09" в часы "пик". В результате перегрузки справочной службы повышается нервно-психическая напряженность труда телефонисток, ухудшается качество обслуживания населения.

В данной работе мы исследовали справочные службы "09" Минска, Москвы и Ленинграда и разработали технические рекомендации по повышению эффективности телефонных разговоров. Нами проведен анализ более 3000 диалогов телефонисток с абонентами и дана инженерно-психологическая оценка автоматизированных рабочих мест в Минске и Ленинграде, где на справочных службах вместо обычных картотек внедряются дисплейные рабочие места с синтезаторами речи.

Анализ речевых диалогов показал, что они содержат много однотипных, часто повторяющихся фраз-реплик. Приблизительно 30% всего объема времени составляют запросы о телефонах служб, пользующихся повышенным спросом у населения: запрашиваются номера телефонов такси, авто- и железнодорожных вокзалов, касс аэрофлота, справочные аптек и службы быта. Такие шаблонные ответы в деятельности телефонисток можно автоматизировать с помощью компьютерного устно-речевого вывода информации через синтезатор речи.

Ниже в таблице представлены типовые запросы абонентов и стандартизированные речевые ответы телефонисток, которые рекомендуются для передачи с помощью синтезатора речи.

затрудняет деятельность телефонистки при наборе и вводе в ЭВМ русских текстов.

2. Отсутствует словарь условных сокращений, используемых при вводе информации в ЭВМ. Нет логической связи между начертанием знака и его значением, смыслом выполненной операции.

3. Важные и часто используемые в работе клавиши расположены бессистемно и "зашумлены" избыточной, лишней клавиатурой.

4. Клавиатуры пульта управления и дисплея слабо взаимосвязаны в пространстве и находятся в различных рабочих плоскостях.

Для устранения указанных недостатков предлагаются следующие рекомендации:

1. На клавиатуре дисплея необходимо устраниить латинский шрифт и оставить только буквы русского алфавита.

2. Все наборное поле целесообразно разделить на две относительно независимые зоны:

- набора буквенної информации (располагается слева);
- выбора команд и управляющих воздействий, а также набора цифровой информации (располагается справа).

3. Необходим словарь условных обозначений с использованием в телефонии сокращений и аббревиатур.

4. Компоновка и взаимное расположение функциональной, буквенной и цифровой клавиатуры должны учитывать частоту и последовательность действий телефонистки.

5. Не рекомендуется жесткая фиксация в пространстве клавиатуры и экрана дисплея. Необходимо предоставить свободу выбора их взаимного расположения в зависимости от индивидуальных желаний телефонисток.

Проведенный анализ диалогов показал, что значительная доля конфликтов приходится на заключительную фазу диалога. Выявлены следующие типовые ошибки ("брэк") в деятельности телефонисток:

I) часто справка-ответ выдается сразу же после запроса абонента, без предварительной паузы и предупреждения. Длительность паузы между вопросом и ответом составляет лишь 1-2 с, и многие абоненты не успевают подготовиться к приему и записи информации;

2) обычно номер телефона телефонистка произносит слитно, без расчленения цифр на пары и без пауз между ними;

3) номер телефона диктуется с ускорением темпа речи и понижением интонации на последних цифрах, что затрудняет их восприятие. 6-значный номер телефона диктуется быстро, за 1,5-1,8 с;

4) после быстрого произнесения номера телефона телефонистка сразу стремится дать "отбой" без получения от абонента подтверждения правильности приема информации.

Такая нерациональная "экономия" времени деятельности телефонисток при выдаче номера телефона создает дополнительную, "параситную" нагрузку на службе "09" из-за повторных звонков извращенных абонентов, которые не успели зафиксировать информацию. Это приводит к возникновению лишних "эмоций" в работе невинных телефонисток, вынужденных обслуживать этих абонентов.

Для того чтобы абоненты успели надежно записать информацию, рекомендуется автоматизировать эту часть диалога с помощью синтезаторов речи. Сокращение времени диалога только на 1с позволяет дополнительно обслужить в Минске 1500, в Ленинграде 5000, а в Москве - 11000 абонентов. По предварительным расчетам автоматизация рутинных речевых действий телефонисток (стандартные ответы в типовых ситуациях, выдача часто запрашиваемой информации и номеров телефонов) позволяет сократить время диалога на 3-5 с. Кроме повышения пропускной способности службы "09" значительно снижается непроизводительная "голосовая" нагрузка телефонисток.

СОДЕРЖАНИЕ

	Стр.
ВВЕДЕНИЕ	3
ЧАСТЬ I. АНАЛИЗ И РАСПОЗНАВАНИЕ РЕЧЕВОГО СИГНАЛА	
Дегтярев Н.П. Параллельно-последовательная модель анализа обнаружения и интерпретации сигналов слитной речи	4
Александровский В.И. Особенности реализации модели анализа, обнаружения и интерпретации слитной речи на базе персонального вычислительного комплекса	25
Рылов А.С. Исследование метода оценки длины речеобразующего тракта по коэффициентам поглощения.....	33
Рылов А.С., Левковская Т.В. Частотно-адаптивный авторегрессионный анализ речевых сигналов...	42
ЧАСТЬ II. СИНТЕЗ РЕЧИ	
Лобанов Б.М. Микроволновой синтез речи по тексту	57
Лосик Г.В. Компьютер с синтезатором речи для слепых	74
Сихеев С.С. Инженерно-психологические особенности использования синтезатора речи на справочной службе "09"	81

АНАЛИЗ И СИНТЕЗ РЕЧИ

Сборник научных трудов

Ответственный за выпуск Н.А.Рудая
Редактор Г.Б.Гончаренко

Подписано к печати 13.02.91.
Формат бумаги 60x84 1/16. Бумага типографская. Офсетная печать.
Уч.-изд.л. 5,6. Усл.печ.л. 5,2. Тираж 300 экз. Зак.58.
Цена 1 р.Юк.

Институт технической кибернетики АН БССР, 220605, г.Минск,
Сурганова, 6
Отпечатано на ротапринте Института технической кибернетики
АН БССР, 220605, Минск, Сурганова, 6

РЕФЕРАТЫ

УДК 621.321

Дегтярев Н.П. Параллельно-последовательная модель анализа, обнаружения и интерпретации сигналов слитной речи//Анализ и синтез речи.- Минск, 1991.- С.4-24.

Анализируются основные компоненты трехуровневой параллельно-последовательной модели анализа, обнаружения и интерпретации речевых сигналов, допускающей решение задачи распознавания и понимания слитной речи многих дикторов в условиях воздействия акустических помех.

Задача повышения помехоустойчивости распознавания решается путем использования принципа обнаружения и классификации в текущем речевом сигнале локальных сегментов, удовлетворяющих критерию достаточного правдоподобия с эталонами заданного словаря. Такой подход, одновременно с повышением помехоустойчивости распознавания, позволяет перейти к реализации механизмов распознавания и понимания слитной речи при условии, что при формировании эталонов учитывается действие звуковых законов слитной речи.

Обосновывается модуляционная природа акустических инвариантов многодикторного описания артикуляции речи. В качестве параметров акустического описания артикуляции речи предлагаются параметры двухформантной модели аппроксимации спектров речи.

Ил.4, библ. 34 назв.

УДК 628.421

Александровский В.И. Особенности реализации модели анализа, обнаружения и интерпретации сигналов слитной речи на базе персонального вычислительного комплекса//Анализ и синтез речи.-Минск, 1991.-С.25-32.

При обсуждении особенностей реализации основных блоков модели анализа, обнаружения и интерпретации сигналов слитной речи на базе персонального вычислительного комплекса показано, что эффективная работа модели может быть обеспечена на пути сочетания специализированных вычислительных блоков и цепей контейнерной обработки данных.

Ил.1, библ.4 назв.

УДК 007.001.362+691.327

Рылов А.С. Исследование метода оценки длины речеобразующего тракта по коэффициентам поглощения//Анализ и синтез речи.- Минск, 1991.-С.33-41.

Решается актуальная задача оценки геометрических параметров речевого тракта диктора по форме речевой волны. В частности, предложен новый метод расчета длины речеобразующего тракта. Этот геометрический параметр может быть использован как критерий качества при создании адаптивных анализаторов для распознавания речи произвольного диктора.

Ил.3, табл.3, библ.6 назв.

УДК 007.001.362+681.327

Рылов А.С., Левковская Т.В. Частотно-адаптивный авторегрессионный анализ речевых сигналов//Анализ и синтез речи.-Минск, 1991.- С. 42-56.

Разработаны алгоритмы адаптивного анализа речи, позволяющие снизить априорную неопределенность речевого сигнала за счет предварительной оценки длины речеобразующего тракта диктора, которая является критерием качества для ширины полосы входного сигнала в данный момент. Результаты экспериментальных исследований описанного метода анализа свидетельствуют о целесообразности его использования для создания распознавающих систем произвольного диктора при условии точной оценки длины речеобразующего тракта данного диктора.

Ил.8, табл.1, библ.9 назв.

УДК 621.391

Лобанов Б.М. Микроволновой синтез речи по тексту//Анализ и синтез речи.-Минск, 1991.-С.57-73.

Описывается новый метод синтеза речи по тексту, отличающийся от формантного метода. Анализируются достоинства и недостатки обоих методов. Даётся подробное описание алгоритма формирования речевой волны из микроволн, алгоритма управления длительностью, высотой тона, амплитудой гласных и согласных звуков. Отдельно рассматривается алгоритм формирования плавно-

гс перехода от сигнала предыдущей фонемы к последующей.
Ил.7, табл.3, библ.4 назв.

УДК 621.320

Лосик Г.В. Компьютер с синтезатором речи для слепых //Анализ и синтез речи.- Минск, 1991.-С.74-80.

Статья посвящена проблеме программно-технического создания компьютера для инвалидов по зрению. Рассматривается круг задач, которые незрячий может решить с помощью ПЭВМ, техника вывода информации из ПЭВМ через синтезатор, эргономика управления выводом, переспросом, темпом.

Рассматривается состав ПП для АРМа слепого человека: "Калькулятор", "Клавиатура", "Записная книжка", "Часы".

УДК 621.352

Сихеев С.С. Инженерно-психологические особенности использования синтезатора речи на справочной службе "09"//Анализ и синтез речи.- Минск, 1991.-С.81-85.

На материале исследования телефонных справочных служб Минска, Москвы, Ленинграда делаются выводы о техническом усовершенствовании организации телефонных разговоров. Дается рекомендации о использовании автоматизированного речевого ответа через синтезатор речи. Приводится перечень речевых сообщений, которые удобно перевести на синтезатор речи в первую очередь.

Табл.1.