УДК 004.8

Ю.В. КРАК, Ю.Г. КРИВОНОС, А.И. КУЛЯС

ПРИКЛАДНЫЕ АСПЕКТЫ СИНТЕЗА И АНАЛИЗА РЕЧЕВОЙ ИНФОРМАЦИИ

Ключевые слова: *синтез, конкатенация, речевой сигнал, автоматизация документирования.*

ВВЕДЕНИЕ

Подсистемы компьютерного синтеза голосовой информации являются неотъемлемой составной частью современного человеко-машинного интерфейса. Они повсеместно используются в высокоинтеллектуальных мультимедийных технологиях, при создании учебных программ, виртуальных сред, библиотечных и других справочников, в web-системах и телекоммуникационных системах IP-телефонии, приложениях для людей со специальными потребностями и т.п. В настоящее время во всем мире, в и частности в Украине, активно исследуются и успешно решаются проблемы синтеза и распознавания голосовых языковых сигналов, моделирования речевого аппарата человека, автоматизации компьютерного документирования аудиоинформации. Синтез естественных языков — важный функциональный компонент систем искусственного интеллекта, поскольку допускает обычный и удобный для человека способ общения. В данном направлении исследований актуальна разработка методов синтеза с максимальным использованием характеристик речи и учетом просодии и интонационных свойств естественного языка.

Значимой частью исследования в области получения искусственной голосовой информации является создание компьютерного артикуляторного синтезатора национальных языков с применением математических методов моделирования звука. При таком подходе необходимо совместно использовать физические модели голосового источника и речевого аппарата человека и разрабатывать математические методы и численные алгоритмы для решения задач акустики.

Автоматизация документирования информации, в частности расшифровка стенограмм заседаний, — необходимая составляющая в работе многих организаций. Как правило, процесс создания и расшифровки стенограмм достаточно продолжителен и попытки его ускорения путем увеличения числа персонала малоэффективны. Для решения таких проблем следует разрабатывать методы предварительной подготовки исходной информации (разбиение на сегменты, очистка от шумов, повышение уровня звучания), на основании которых создавать системы распределенного компьютерного документирования.

Цель настоящей работы — исследование и разработка новых математических методов и усовершенствование существующих подходов к практическому решению проблем автоматизации синтеза и анализа речевой информации.

КОНКАТЕНАТИВНЫЙ СИНТЕЗ РЕЧЕВОЙ ИНФОРМАЦИИ

Системы синтеза речи можно классифицировать по способам получения голосового сигнала [1, 2]. Выделяют три основных метода синтеза: артикуляционный, формантный и конкатенативный. В системах конкатенативного синтеза процесс построения выходного акустического сигнала базируется на основе последовательного объединения необходимых элементов синтеза. Основной целью синтеза естественной речи является построение алгоритмов озвучивания информации с наибольшим приближением характеристик звучания к голосу человека. Процесс конкатенации определяется структурой и наполнением базы данных элементов синтеза, поэтому с повышением качества синтеза

растет и размерность элементной базы для него. Существует несколько стандартных подходов к выбору концепции формирования минимальных элементов синтеза — фонов, аллофонов, дифонов, слогов, фонем-трифонов и т.п. Естественность и качество звучания синтезированных речевых сигналов объясняется тем, что в системах конкатенативного синтеза как элементы синтеза используются природные речевые сигналы, записанные профессиональными дикторами — носителями языка. При выборе элементов синтеза необходимо учитывать большое число фонетических коартикуляционных свойств языка, таких как: акустические особенности и характеристики изолированных звуков и звуков в слове; просодия и интонационные свойства естественного языка в зависимости от расположения элементов в слове, а также смягчение, удвоение; влияние акцентированности и ударения на частотные характеристики речи; суперсегментные явления и характеристики частоты основного тона; структуризация пары «объект синтеза–элемент синтеза» и др.

В зависимости от элементов синтеза, которые используются при генерировании сигналов озвучивания текстовой информации, выделяются такие методы конкатенативного синтеза: аллофонный, дифонный, слоговой, трифонный, сегментивный. В настоящем исследовании используем сегментивный синтез, при котором слово представляется через три части: префиксную (начальную), внутреннюю и суффиксную (конечную) [3]. Во множествах этих сегментов содержатся сегменты как с ударениями, так и безударные, а слово подается последовательностью определенных сегментов, уже включающих просодию. Из соответствующего множества сегментов выбираются те, которые состоят из наибольшего количества фонем. Это уменьшает количество конкатенаций сегментов и повышает естественность звучания синтезированной речи.

Математическую модель представления текстовой информации для конкатенативного сегментивного синтеза речи запишем следующим образом:

$$w_l = s_j \{\mathrm{i}\} f_k. \qquad (1)$$

Здесь $w_l \in W \ \forall l \in N$ — слово из множества слов некоторого языка W; $s_j \in S$ $\forall j \in N$ — начальный сегмент из множества префиксных сегментов S; $\{\mathrm{i}\}$ — некоторый (определенный для каждого конкретного слова) последовательный набор внутренних сегментов $i_m \in I \ \forall m \in N$ из множества внутренних сегментов I; $f_k \in F \ \forall k \in N$ — сегмент из множества суффиксных сегментов F; $s_j, i_m, f_k \in \Omega \ \forall j, m, k \in N$, Ω — множество всех сегментов языка.

Представление в виде модели (1) позволяет эффективно структурировать и существенно повысить скорость работы систем речевого синтеза благодаря простоте получения и обработки элементов синтеза. Таким образом, основные усилия в области разработки систем синтеза естественной речи без модификации элементов речи направлены на совершенствование баз элементов синтеза и алгоритмов сегментации для улучшения характеристик звучания и просодического оформления синтезированной речи. Часть элементов синтеза может обладать общими для различных языков фонемными и просодическими характеристиками, но разрабатывать мультиязычные системы синтеза со звучанием, аналогичным естественному, можно лишь при использовании современных математических методов и алгоритмов модификации входных элементов синтеза, например средствами модификации периодов основного тона [4].

В модели (1) минимальным семантически полным носителем информации является слово. Это подтверждается результатами фонетических исследований, которые показали зависимость просодических характеристик звучания речи от их расположения в слове. Слово, в рамках структурной пары «объект синтеза–элемент синтеза», представлено схемой на рис. 1.

Как следует из данной схемы, объект синтеза (слово) представляется последовательностью из множества элементов синтеза $\{SE\}$. Каждый элемент синтеза SE имеет следующую структуру: текстовое представление элемента синтеза

$Text(SE) \in \Omega$; список слов, содержащих элемент синтеза $WList(SE) = \{w_i\}$, где $i \in N$, $w \in W$; рейтинг, т.е. количество слов, в которых содержится элемент синтеза $R(SE)$; звуковое представление элемента синтеза $Audio(SE)$. При этом каждый объект синтеза SO имеет такую структуру: текстовое представление объекта синтеза $Text(SO) \in W$; последовательный список элементов синтеза слова $WList(SO) = s_j \{\mathrm{i}\} f_k$, $s_j \in S \ \forall j \in N$, $\{i : i_m \in I \ \ \forall m \in N\}$, $s_j, i_m, f_k \in \Omega \ \forall j, m, k \in N$; рейтинг, т.е. число элементов синтеза, которые составляют слово $R(SO)$.

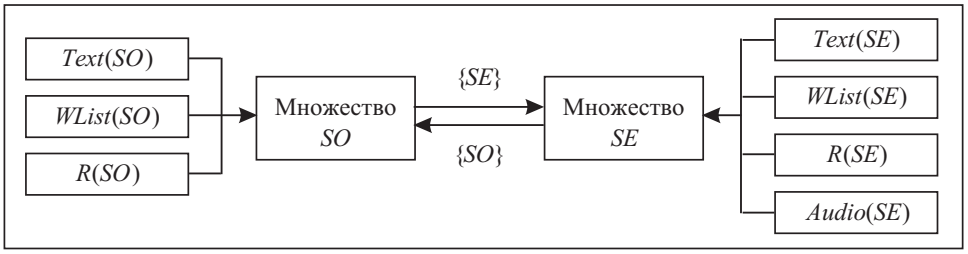

Рис. 1. Схема объектно-элементной модели представления информации

Таким образом, качество синтезированной речи, построенной на принципах конкатенации, непосредственно зависит от языковых баз данных. Чем больше объем голосовых сегментов базы данных и, следовательно, полнее представлена в ней звуковая, интонационная, темпоральная вариативность речи, тем большую естественность синтезированного звучания можно получить [5].

Рассмотрим вопросы создания баз элементов синтеза для славянских языков, в частности украинского. Исследование речи и анализ свойств ее звучания показали необходимость учета большого количества взаимосвязанных факторов, влияющих на формирование элементов синтеза. Следует учитывать позиционную обусловленность элементов синтеза, ударяемость гласных звуков, смягчение и удвоенность согласных звуков, свойства коартикуляционных переходов между согласными и гласными звуками, раздельность звучания согласных и гласных звуков, совпадение согласных звуков и т.п. Исходя из результатов фонетических исследований украинской речи, можно получить следующие признаки естественности ее звучания: G_1 — влияние согласных звуков на гласные (данный признак определяет связность произношения звуков, входящих в состав слов и словосочетаний, взаимное приспособление артикуляции звуков за счет наложения конечной и начальной фаз соседних звуков, выделения переходных и основной частей гласных, лабиализацию согласных перед звуками [у] и [о], а также предопределяет использование слогов в качестве основных сегментов именно через коартикуляционные переходы между фонемами); G_2 — твердость или мягкость согласных (в зависимости от смягчения согласного в слове его звучание изменяется); G_3 — раздельность звучания согласных и гласных звуков, например открытые слоги типа «согласная–апостроф–гласная»; G_4 — удвоение и удлинение согласных (удлиненные согласные имеют характерное звучание, искусственное создание которого — сложная и неоднозначная задача из коартикуляционной модификации согласных); G_5 — упрощение согласных при произношении (наиболее часто упрощения происходят в суффиксах слов при совпадении нескольких согласных); G_6 — отличие звучания ударяемых и безударных гласных (приводит к необходимости создания как ударяемых, так и безударных сегментов); G_p — позиционность сегмента (характеризует расположение сегмента в слове — в начале, середине или конце).

С учетом важной роли позиционности сегментов в слове, основными признаками являются G_1, G_6, G_p, формирующие свойства естественности звучания синтезированной речи. Признаки $G_1 - G_5$ определяют свойства и взаимосвязи звучания фонем и таким образом формируют мультифонемность звучания. Схема взаимосвязи признаков приведена на рис. 2, где G_s — признак ударного гласного, а G_m — совокупность свойств признаков $G_1 - G_5$.

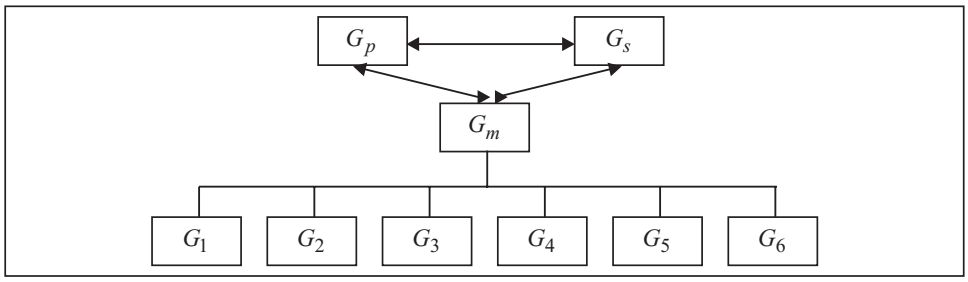

Рис. 2. Схема взаимосвязи признаков естественности звучания речи для конкатенативного сегментивного синтеза

Как следует из схемы, для естественности звучания каждый сегмент одновременно должен определяться следующими признаками: позиционность, ударение и мультифонемность.

Практическая реализация данного подхода к синтезу украинской речи показала, что для создания качественного синтезатора необходимо записать и соответствующим образом обработать около 5000 сегментов, что является вполне достижимой задачей. Тестовые варианты такого сегментивного синтеза показали его эффективность и продемонстрировали естественность звучания синтезированной речи.

ИССЛЕДОВАНИЕ ГОЛОСОВОЙ ИНФОРМАЦИИ НА ОСНОВАНИИ ФИЗИЧЕСКИХ МОДЕЛЕЙ

Важной проблемой для анализа и синтеза речевой информации является компьютерное воспроизведение звуков, произносимых человеком в процессе разговора, с использованием физических моделей голосового источника и речевого тракта. С этой целью разработаны численные алгоритмы и программное обеспечение для решения математических задач, описывающие голосовой источник и речевой тракт человека, а также численные алгоритмы для решения обратной задачи для речевого тракта. Прикладное значение таких исследований заключается в создании численных алгоритмов для общего моделирования голосовых связок и речевого тракта, которые могут использоваться для построения компьютерного артикуляторного синтезатора речи [6].

Модель голосовых связок [7] описывает колебание двух масс, связанных пружинами как со стенками тракта, так и между собой. Допускается, что связки являются двусторонне симметричными. Массы, моделирующие связки, осуществляют колебание в поперечном к движению воздуха направлении. Система уравнений для двух масс, которые колеблются, записывается в виде

$$\begin{aligned} m_1 \frac{d^2 x_1(t)}{dt^2} + r_1 \frac{dx_1(t)}{dt} + k_1(x_1(t) - x_{01}) + k_c(x_1(t) - x_2(t)) = l_g d_1 p_{m_1}(t), \\ m_2 \frac{d^2 x_2(t)}{dt^2} + r_2 \frac{dx_2(t)}{dt} + k_2(x_2(t) - x_{02}) - k_c(x_1(t) - x_2(t)) = l_g d_2 p_{m_2}(t), \end{aligned} \tag{2}$$

где m_1 и m_2 — массы; $x_1(t)$, $x_2(t)$ — смещение масс m_1 и m_2; x_{01}, x_{02} — начальное положение масс m_1 и m_2; t — время; r_1 , r_2 — коэффициенты демпфирования; k_1, k_2 — упругость пружин для масс m_1 и m_2; k_c — упругость пружины, соединяющей массы m_1 и m_2; d_1, d_2 — толщина масс m_1 и m_2; l_g — действующая длина голосовых связок; $l_g d_1$, $l_g d_2$ — поверхности масс m_1 и m_2, на которые действует давление $p_{m_1}(t)$ и $p_{m_2}(t)$ соответственно.

Распределение давления в голосовой щели аппроксимируется последовательными дискретными значениями p_{ij} на каждом *j*-м конце каждой *i*-й массы, откуда получаем систему уравнений для изменений давления в виде

$$p_s - p_{11}(t) = 0{,}69\rho \frac{u_g^2(t)}{A_{g_1}^2(t)} + \int_0^c \frac{\rho}{A_c(x)}\, dx\, \frac{du_g}{dt},$$

$$p_{11}(t) - p_{12}(t) = 12\nu d_1 \frac{l_g^2 u_g(t)}{A_{g_1}^3(t)} + \frac{\rho d_1}{A_{g_1}} \frac{du_g}{dt},$$

$$p_{12}(t) - p_{21}(t) = \frac{1}{2}\rho\, u_g^2(t) \left(\frac{1}{A_{g_2}^2(t)} - \frac{1}{A_{g_1}^2(t)} \right),$$

$$p_{21}(t) - p_{22}(t) = 12\nu d_2 \frac{l_g^2 u_g(t)}{A_{g_2}^3(t)} + \frac{\rho d_2}{A_{g_2}} \frac{du_g}{dt},$$

$$p_{22}(t) - p = \frac{1}{2}\rho \frac{u_g^2(t)}{A_{g_2}^2(t)} \left[2 \frac{A_{g_2}(t)}{A_1} \left(1 - \frac{A_{g_2}(t)}{A_1} \right) \right], \qquad (3)$$

где p_s — давление на входе в голосовую щель; p — атмосферное давление; ρ — плотность воздуха; ν — сдвиговая вязкость воздуха; A_1 — площадь голосового тракта на входе; A_{g_i} — площадь голосовой щели под i-й массой; $u_g(t)$ — поток воздуха; $A_{g_i}(t) = (A_{g_{0i}} + 2l_g x_i(t))$, $i = 1, 2$ $(x_1(t) \ge x_{01}$, $x_2(t) \ge x_{02}$, $A_{g_{01}}$, $A_{g_{02}}$ — остаточные площади в момент соединения голосовых связок).

Для определения значений давления $p_{m_1}(t)$, $p_{m_2}(t)$ используются следующие соотношения:

$$p_{m_1}(t) = \frac{1}{2}(p_{11}(t) + p_{12}(t)), \quad p_{m_2}(t) = \frac{1}{2}(p_{21}(t) + p_{22}(t)). \qquad (4)$$

Решением системы (2)–(4) является функция $u_g(t)$, определяющая поток воздуха на выходе из голосовой щели.

Для реализации двухмассовой модели голосовых связок (2)–(4) разработан численный метод, в основу которого положена комбинация метода решения системы уравнений колебания двух масс и метода решения нелинейной системы для изменений давления. Для проверки адекватности результатов моделирования проведена серия численных экспериментов, которые показали, что значение исходных параметров модели и рассчитанные характеристики голосового источника находятся в физиологически допустимых пределах [8].

Полученное с помощью разработанного численного алгоритма решение и его производная используются в качестве голосового источника для моделей речевого тракта. Для моделирования распространения акустических волн в речевом тракте человека как в неоднородной акустической трубе, которая начинается между голосовыми связками и заканчивается губами, применена система уравнений акустики в частных производных

$$-S(x)\frac{\partial p}{\partial x} = \rho \frac{\partial u}{\partial t},$$

$$-\rho c^2 \frac{\partial u}{\partial x} = S(x)\frac{\partial p}{\partial t}, \qquad (5)$$

где $0 \le x \le L$, $t > 0$, L — длина речевого тракта; $p(x,t)$ — давление в тракте в момент времени t; $u(x,t)$ — объемная скорость потока; ρ — плотность воздуха в тракте; c — скорость звука; $S(x)$ — функция площади поперечного сечения.

Поскольку тракт имеет неоднородное поперечное сечение, он разбивается на цилиндрические секции одинаковой длины с постоянной площадью сечения. В качестве краевого условия на входе в тракт принимается поток $u_g(t)$ из (2)–(4). Исходя из этого имеем для системы (5) краевое условие $u(0,t) = u_g(t)$. На противоположном конце тракта задано условие $p(L,t) = 0$. Разностная задача для аппроксимации системы уравнений (5) построена на разнесенной сетке. Для решения применяется метод «чехарда».

Для моделирования распространения акустических волн в тракте также использовано уравнение Вебстера

$$S(x)\frac{\partial^2 P}{\partial t^2} = c^2\,\frac{\partial}{\partial x}\left(S(x)\frac{\partial P}{\partial x}\right), \tag{6}$$

где x — пространственная координата вдоль средней линии тракта в среднесагитальной плоскости; t — момент времени; $p(x,t)$ — искомое давление в тракте; $S(x)$ — профиль площадей поперечного сечения вдоль тракта.

В качестве краевого условия на входе в тракт взята производная от потока воздуха $P(0,t) = -\dfrac{\rho}{S(0)}\dfrac{du_g(t)}{dt}$. Для решения задачи (6) применен конечно-разностный метод. Для решения системы разностных уравнений использован итерационный метод последовательной верхней релаксации.

Рассмотрена задача восстановления формы речевого тракта по измеренным акустическим параметрам сигнала на базе акустического уравнения Клейна–Гордона [9]. Для этого введена новая переменная, определенная выражением

$$\varphi(x,t) = P(x,t)S(x)^{1/2}, \tag{7}$$

что позволило записать акустическое уравнение в форме Клейна–Гордона

$$\frac{\partial^2\varphi(x,t)}{\partial t^2} = c^2\,\frac{\partial^2\varphi(x,t)}{\partial x^2} - c^2U(x)\varphi(x,t),\ \ 0< x< L,\ \ 0< t\le T. \tag{8}$$

Уравнение (8) имеет форму волнового, функция $U(x)$ определена в терминах площади поперечного сечения речевого тракта как

$$U(x) = \frac{d^2S(x)^{1/2}\,/\,dx^2}{S(x)^{1/2}}. \tag{9}$$

Обратная речевая задача — это задача нахождения функции $S(x)$ по измеренным параметрам речевого сигнала на выходе из тракта. Математически данная задача решается как задача поиска минимума некоторого функционала при разного рода ограничениях. Пусть на выходе из тракта измеряется давление $P(L,t)$, связанное с решением уравнения Клейна–Гордона соотношением (7). Обозначим $\Phi(t)$ функцию, которая измеряется на выходе из тракта. Тогда обратная речевая задача сводится к минимизации функционала

$$J(U) = \int_0^T (\Phi(t) - \varphi_U(L,t))^2\,dt, \tag{10}$$

где $\varphi_U(L,t)$ — решение задачи (8) при заданной функции $U(x)$.

Для минимизации функционала (10) использован градиентный метод. Прирост функционалу запишем в виде

$$\Delta J(U) = J(U+h) - J(U) = \int_0^T 2(\Phi(t) - \varphi_U(L,t))\Delta\varphi\,dt + \int_0^T (\Delta\varphi)^2\,dt,$$

где $\Delta\varphi = \varphi_{U+h}(x,t) - \varphi_U(x,t)$. Для определения градиента функционала сопряженная задача записывается в виде

$$\frac{\partial^2 \Psi(x,t)}{\partial t^2} = c^2 \frac{\partial^2 \Psi(x,t)}{\partial x^2} - c^2 U(x)\Psi(x,t),\ 0 < x < L,\ 0 < t \le T.$$

Градиент функционала определяется через решение сопряженной задачи по формуле $J' = -\varphi\Psi$.

После определения $U(x)$ функция $S(x)$ находится из (9).

В тестовых расчетах использовалась следующая стратегия проверки работоспособности построенного алгоритма и созданного программного обеспечения. Решалась прямая задача и определялся сигнал на выходе. Далее для решения обратной задачи этот сигнал использовался в качестве начального приближения. Точность решения оценивалась по процедуре ресинтеза: синтезированный по найденному решению сигнал не должен значительно отличаться от исходного сигнала, по параметрам которого решалась обратная задача. Результаты численного моделирования продемонстрировали адекватность и эффективность разработанного подхода к решению обратной речевой задачи.

АНАЛИЗ ГОЛОСОВЫХ СИГНАЛОВ ДЛЯ РЕШЕНИЯ ПРОБЛЕМ АВТОМАТИЗАЦИИ ДОКУМЕНТИРОВАНИЯ ИНФОРМАЦИИ

Одной из важных задач при создании систем автоматизации компьютерного документирования речевой информации является разделение входящего сигнала на равноценные сегменты. Для сегментации сигнала используем информацию о паузах, имеющихся в сигнале, а также позициях в звуковом сигнале, в которых происходит смена диктора. Поиск пауз в сигнале будем осуществлять путем сравнения энергии в анализируемом фрейме с некоторым пороговым значением. Для этого проходим по сигналу прямоугольным окном продолжительностью $\tau = 50$ мс таким образом, чтобы начало каждого последующего окна приходилось на середину предыдущего. Полагаем, что на участке сигнала длиной 10 с должна быть по крайней мере одна пауза. Считаем, что уровень энергии сигнала на участках, соответствующих паузам, ниже, чем на участках, где существует голосовая активность. Энергию E_{s_k} сигнала $s[i]$ на интервале 10 мс определим как дисперсию амплитуды сигнала в заданном окне k:

$$E_{s_k} = \log_{10}\left(\frac{1}{N}\sum_{i=1}^{N} s_k[i]^2 - \left(\frac{1}{N}\sum_{i=1}^{N} s_k[i]\right)^2\right). \qquad (11)$$

Здесь $s_k[i] = s[i]$ для $\tau \times D \times k \le i < 3\tau \times D \times k$ и $s_k[i] = 0$ в других случаях; D — частота дискретизации сигнала.

К полученным уровням энергии (11) применим метод медианного сглаживания пятого порядка для устранения негативного влияния случайных возмущений на измерения. Для принятия решения о том, соответствует ли анализируемый фрейм паузе, значение энергии в нем сравнивается с порогом. Поскольку условия и уровень шума в сигнале могут меняться со временем, необходимо динамически рассчитывать порог в процессе обработки сигнала. Предложен следующий алгоритм адаптивного вычисления порога энергии для пауз. На участке сигнала временной протяженностью 10 с, предшествующем анализируемому фрейму, находятся минимальный и максимальный уровни энергии для данного участка: $E_{\min}$ и $E_{\max}$. Далее уровень энергии E в текущем фрейме сравнивается с полученными значениями. Решение о принадлежности текущего фрейма паузе принимается, если выполняется условие

$$E < E_{\min} \vee \frac{E - E_{\min}}{E_{\max} - E_{\min}} < 0{,}2\,. \qquad (12)$$

Значения $E_{\min}$ и $E_{\max}$ уточняются на каждом шаге алгоритма с учетом предыдущих 10 с звучания. Фреймы с низким уровнем энергии, расположенные последовательно один за другим, объединяются в одну паузу. Паузы, длина которых меньше некоторой заданной длины, исключаются из рассмотрения, так как они наиболее вероятно соответствуют участкам с низкой энергией в середине слова (например, шипящим согласным).

При сегментации сигнала целесообразно учитывать позиции в сигнале, где происходит смена диктора [10]. Полагаем, что смена диктора в сигнале происходит в промежутке паузы, т.е. после того, как заканчивает говорить один диктор и начинает говорить другой, возникает пауза. На практике это не всегда так, дикторы могут перебивать друг друга. Однако такие ситуации сложно учитывать при сегментации сигнала и в данной работе они не рассматриваются.

Полагаем, что $X = \{x_1, x_2, \ldots, x_n\}$ и $Y = \{y_1, y_2, \ldots, y_n\}$ — множества характеристических векторов, определенных на участке сигнала соответственно до паузы и между текущей и последующей паузой, а N_x и N_y — количество точек в первом и втором множествах. Характеристические векторы в данном случае представляют собой четырнадцатимерные векторы, в которых 13 мел-кепстральных коэффициентов, рассчитанных на участке сигнала продолжительностью 30 мс, и задана частота основного тона.

Пусть $Z = X \cup Y$ — объединение множеств характеристических векторов с количеством точек $N = N_x + N_y$. Множества X и Y сравниваются с помощью некоторой меры различия, и если они значительно отличаются, принимается решение о том, что в анализируемом участке сигнала происходит смена диктора.

Задачу определения замены диктора можно сформулировать в виде задачи проверки гипотезы. Пусть H_0 — гипотеза о том, что смены диктора не происходит, а H_1 — гипотеза о том, что замена происходит. Положим также, что векторы, из которых состоят множества X и Y, независимы и имеют одинаковое распределение случайными величинами. Пусть Θ_Z — параметры распределения для множества Z, рассчитанные методом максимального правдоподобия. В таком случае логарифмическое соотношение правдоподобия для множества наблюдений Z при условии выполнения гипотезы H_0 запишется как

$$L_0 = \sum_{i=1}^{N_x} \log\, p(x_i|\Theta_Z) + \sum_{i=1}^{N_y} \log\, p(y_i|\Theta_Z), \tag{13}$$

где $p(x|\Theta)$ — вероятность того, что x выполняется при условии Θ.

Для проверки гипотезы H_1 рассчитываются параметры индивидуальных распределений для наборов наблюдений X и Y, соответственно обозначенные как Θ_X и Θ_Y. Логарифмическое отношение правдоподобия для гипотезы H_1 запишем в виде

$$L_1 = \sum_{i=1}^{N_x} \log\, p(x_i|\Theta_X) + \sum_{i=1}^{N_y} \log\, p(y_i|\Theta_Y). \tag{14}$$

Меру различия для множеств X и Y в таком случае можно задать как байесовский информационный критерий

$$d_1 = L_1 - L_0 - \frac{\lambda}{2}\Delta K \log N. \tag{15}$$

Здесь $\Delta K = N_x - N_y$, а λ — параметр, определенный экспериментально. Решение о смене диктора в анализируемом участке сигнала принимается, если заданная таким образом мера различия превышает некоторый порог, установленный экспериментально.

В целях повышения качества сегментов для их дальнейшей автоматизированной расшифровки осуществляется предварительная цифровая обработка сигнала для уменьшения уровня шума и изменения скорости воспроизведения звукового сигнала без изменения тембра голоса говорящего.

Шумы в сигналах, подаваемых на вход системы компьютерного документирования, могут считаться аддитивными в спектральной области. Следовательно, для фильтрации таких шумов можно применить методы спектрального вычитания или винеровской фильтрации. Для их работы необходимо, чтобы был известен спектр шума. Предположим, что участки сигнала, соответствующие паузам, содержат только шум. Тогда для аппроксимации шума можно использовать паузы, которые были определены выше.

Для изменения скорости воспроизведения сигнала с сохранением тембра голоса диктора следует убедиться, что продолжительность сигнала изменяется, но при этом частота основного тона говорящего должна сохраняться. Обеспечить это возможно с помощью алгоритмов типа PSOLA [4]. Для их реализации вначале решается задача обнаружения периодов псевдопериодичности в звуковом сигнале (питч-периодов). Для этого исходный звуковой сигнал пропускается через низкочастотный и высокочастотный фильтры с конечными импульсными характеристиками. Далее для сглаживания сигнала каждый элемент вектора исходного сигнала заменяется взвешенным средним четырех окружающих его элементов по формуле

$$d[i] = \frac{3x[i-2] + x[i-1] - x[i+1] - 3x[i+2]}{10}. \tag{16}$$

К полученному сигналу применяется медианное сглаживание порядка $n = 199$ (каждый элемент вектора заменяется медианой вектора, состоящего из n элементов, окружающих текущий элемент). После этого в сигнале обнаруживаются точки, где последовательность, состоящая из элементов вектора сигнала, изменяет знак – на знак +, и такие точки обозначаются как границы питч-периодов. Среди определенных таким образом границ находятся и исключаются точки, расположенные слишком близко одна к другой, а для участков сигнала, где нет явной псевдопериодичности, задаются условные границы с некоторым постоянным интервалом.

После определения границ периодов псевдопериодичности можно изменять акустические характеристики сигнала. Исходный сигнал представим в виде функции периодов основного тона $x_i[n]$:

$$x[n] = \sum_{i=-\infty}^{\infty} x_i[n - t_a[i]], \tag{17}$$

где $t_a[i]$ — границы периодов псевдопериодичности сигнала, т.е. разность между двумя соседними границами $P_a[i] = t_a[i] - t_a[i-1]$ равняется периоду основного тона в момент времени $t_a[i]$. Питч-период определим через исходный сигнал, умноженный на оконную функцию $x_i[n] = w_i[n]x[n]$, где окна w_i удовлетворяют условию $\sum_{i=-\infty}^{\infty} w_i[n - t_a[i]] = 1$, что достигается использованием оконных функций типа Хэннинга или трапециевидным окном длиной в два периода основного тона.

В результате работы алгоритма необходимо получить сигнал $y[n]$, который имеет одинаковые с $x[n]$ спектральные характеристики, но отличается от него основным тоном и/или продолжительностью. Чтобы достичь этого, заменяем аналитические границы питч-периодов $t_a[i]$ границами $t_b[i]$, а аналитические периоды основного тона $x_i[n]$ — периодами $y_i[n]$ согласно формуле

$$y[n] = \sum_{j=-\infty}^{\infty} y_j[n - t_b[j]]. \tag{18}$$

Таким образом, достаточно задать границы $t_b[i]$, соответствующие продолжительности и основному тону, чтобы получить необходимый результат. Период основного тона найдем подстановкой ближайшего периода $y_i[n]$, соответствующего аналитическому периоду $x_i[n]$.

Для проверки эффективности работы предложенного подхода был проведен эксперимент по расшифровке записи защиты диссертации. Продолжительность записи — около двух часов. Время ее преобразования в текст с применением стандартных средств занимает от 12 до 18 часов. Использование предложенной системы позволило сократить это время до четырех часов, а группе стенографистов из пяти человек для обработки записи понадобилось около 40 минут.

ЗАКЛЮЧЕНИЕ

Предложенные в настоящей работе подходы к анализу и синтезу речевой информации показали эффективность и практическую реализуемость. Метод конкатенативного сегментивного синтеза позволяет построить синтез речевой информации с учетом просодических и интонационных характеристик звучания, тем самым приблизив искусственную речь к природной, обычной для человеческого общения.

Приведены новые результаты решения задачи компьютерного моделирования речевых сигналов человека на основании совместного использования моделей голосового источника и речевого тракта человека. Сформулирована и решена обратная задача восстановления параметров речевого тракта по замерянным выходным сигналам.

Рассмотрены новые методы предварительной подготовки речевого сигнала для решения задач автоматизации документирования. Разработана распределенная система компьютерного документирования информации [11]. Проведенные эксперименты продемонстрировали эффективность и перспективность предложенного подхода. Дальнейшие исследования будут направлены на совершенствование методов анализа и синтеза голосовой информации.

СПИСОК ЛИТЕРАТУРЫ

1. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. — Киев: Наук. думка, 1987. — 261 с.
2. Лобанов Б.М., Цирульник Л.И. Компьютерный синтез и клонирование речи. — Минск: Белорус. наука, 2008. — 342 с.
3. Кривонос Ю.Г., Крак Ю.В., Шатковский Н.Н. Структура, свойства, характеристики объектов и элементов синтеза речи // Компьютер. математика. — 2006. — № 1. — С. 61–69.
4. Dutoit T., Leich H. MBR-PSOLA: Text-to-speech synthesis based on an MBE resynthesis of segments database // Speech Communication. — 1993. — N 13. — P. 435–440.
5. Fujisaki H. Prosody, models and spontameous speech // Computing prosody. — New York: Springer-Verlag, 1996. — P. 27–42.
6. Крак Ю.В., Стеля І.О. Чисельне моделювання голосових зв'язок за двомасовою моделлю // Журн. обчисл. та прикл. математики. — 2007. — № 94. — С. 55–60.
7. Ishizaka K., Flanagan J.L. Synthesis of voiced sounds from a two-mass model of vocal cords // Bell System Techn. J. — 1972. — **51**, N 6. — P. 1233–1268.
8. Кривонос Ю.Г., Крак Ю.В., Стеля І.О. Пряма і обернена задачі моделювання мовного апарату людини // Доп. НАН України. — 2011. — № 10. — С. 44–47.
9. Forbes B.J., Pike E.R., Sharp D.B. The accoustical Klein–Gordon equation: The wave-mechanical step and barrier potential functions // J. Acous. Soc. Amer. — 2003. — **114**, N 3. — P. 1291–1302.
10. Крак Ю.В., Куляс А.И., Загваздин А.С. Определение изменения диктора в речевом сигнале // ИИ-2010. Материалы междунар. науч.-техн. конф. 20–24 сент. 2010, Кацивели,АР Крым. — Донецк: ИПШИ «Наука і освіта», 2010. — С. 197–201.
11. Кривонос Ю.Г., Крак Ю.В., Бармак А.В., Загваздин А.С. Информационная система распределенного компьютерного документирования речевых фонограмм заседаний // Управляющие системы и машины. — 2008. — № 3. — С. 46–52.

Поступила 24.01.2013