РЕЧЕВОЙ ИНТЕРФЕЙС ВИРТУАЛЬНОГО СОБЕСЕДНИКА

Б. Лобанов, А. Давыдов, Д. Жадинец, В. Киселёв, Л. Цирульник

Объединённый институт проблем информатики НАН Беларуси

Описывается прототип и разработанная демо-версия виртуального устно-речевого собеседника. Показаны основные этапы разработки и реализации модуля автоматического распознавания ключевых слов в потоке речи, модуля принятия словесных решений, модуля синтеза речи по тексту и модуля менеджера речевого диалога.

Введение

В основу речевого интерфейса для виртуального собеседника положены оригинальные научно-технические решения, полученные сотрудниками Лаборатории распознавания и синтеза речи ОИПИ НАН РБ в течение последних лет [1 – 3]. Уникальность разработанной системы заключается в том, что она позволяет осуществить:

- надёжное распознавание ключевых слов запроса *в непрерывном потоке речи*;
- *многодикторное* распознавание ключевых слов *в условиях акустических помех и искажений;*
- *многоголосый* синтез речи *по произвольному тексту*;
- возможность использования *«клонов» голоса конкретной личности* в процессе синтеза речи;
- реализацию *дуплексного режима в реальном времени* (возможность распознавания ключевых слов одновременно с синтезом речевого ответа).

Общая структура системы

Речевой интерфейс интегрирован в состав компьютерной модели виртуального собеседника (система *РЕВИРС),* в которой реализована возможность создания сценариев диалога для разнообразных приложений и осуществления их посредством устно-речевого человеко-машинного общения. Система РЕВИРС имитирует работу центра автоматической телефонной службы, осуществляющей естественно-речевое общение с абонентами по сценариям “прокат автомобилей”, “гороскоп на неделю”, “погода”.

Пользовательский интерфейс системы РЕВИРС показан на рис. 1.

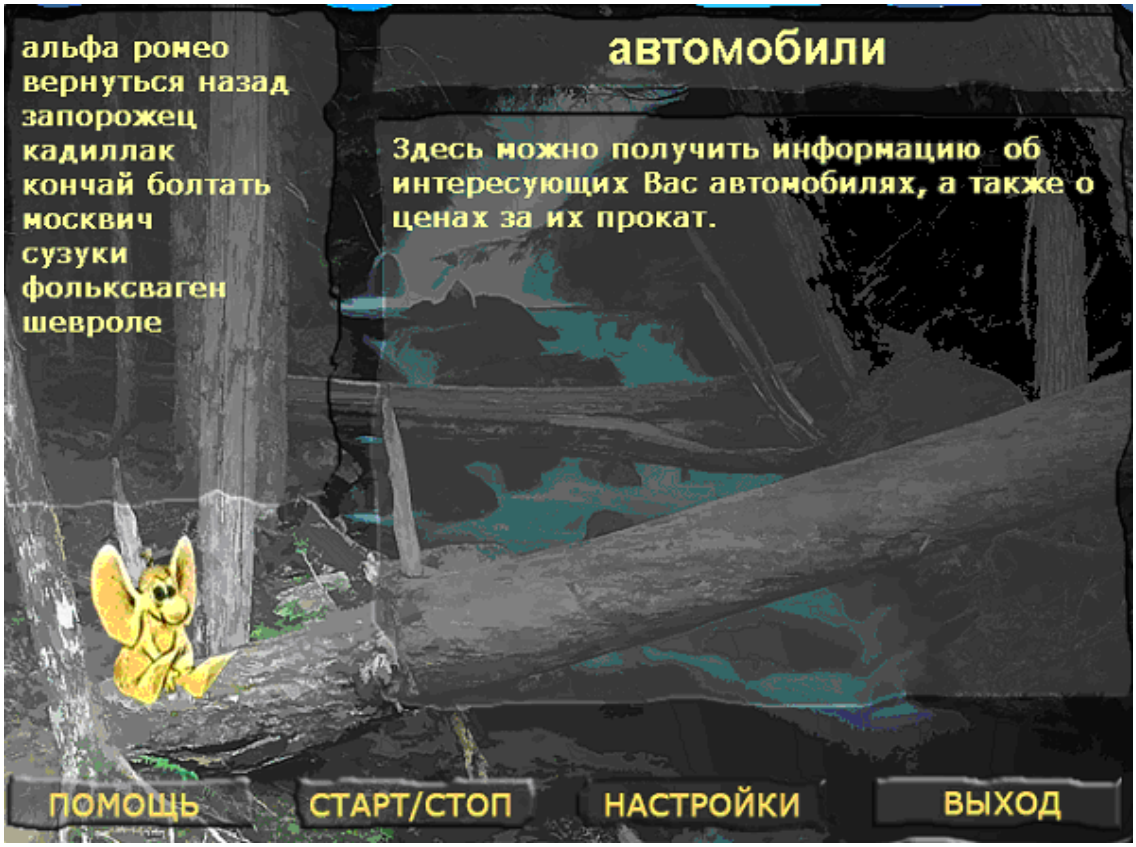

Рис. 1. Интерфейс системы РЕВИРС

Основные функциональные возможности системы:

- распознавание в потоке речи ключевых слов из списка названий сценариев диалогов (“прокат автомобилей”, “гороскоп на неделю”, “погода”);
- распознавание в потоке речи ключевых слов из списка слов по выбранному сценарию (“Кадиллак”, “Шевроле”, и т.д. или “понедельник”, “вторник” и т.д.);
- распознавание в потоке речи ключевых слов из списка общих команд, таких как “вернуться назад”, “кончай болтать ” и т.д.;
- синтез речевых ответов системы по произвольному тексту, задаваемому сценарием;
- выбор голоса для синтеза речи из имеющегося набора «голосовых клонов»;
- использование для ответов как синтезированной, так и естественной (заранее подготовленной и хранящейся в звуковом файле) речи;
- визуальное отображение процесса общения с виртуальным собеседником: ожидание речи, слушание (распознавание), говорение (синтез).

Система предоставляет пользователю возможность изменять существующие или создавать собственные сценарии, изменять и дополнять списки ключевых слов и речевых ответов.

Структурная схема системы РЕВИРС приведена на рис.2.

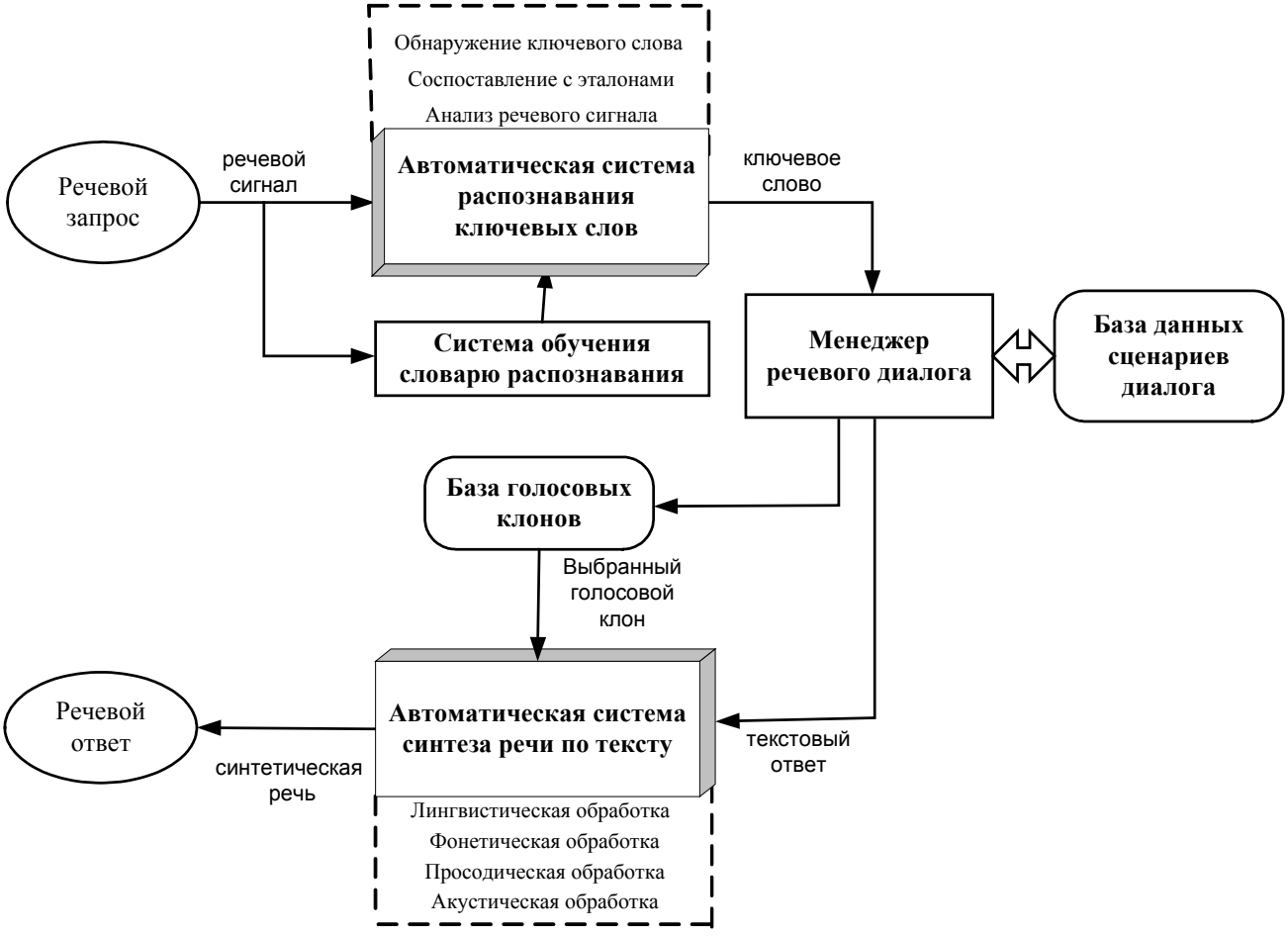

Рис. 2. Структурная схема системы РЕВИРС

Речевой сигнал поступает на вход автоматической системы распознавания ключевых слов, которая осуществляет анализ информативных признаков сигнала, их сопоставление с эталонами ключевых слов, принятие решения об обнаружении ключевого слова. Если ключевое слово обнаружено, оно перенаправляется менеджеру речевого диалога, который формирует текстовый ответ и выбирает голосовой клон для синтеза речевого ответа. Выбранный клон и текст ответа поступают на вход системы синтеза речи, которая осуществляет лингвистическую, фонетическую, просодическую и акустическую обработку, в результате чего текст преобразуется в звучащую речь заданного голосового клона.

Наиболее важные функциональные блоки системы: подсистема синтеза речи по тексту и

подсистема распознавания ключевых слов в потоке речи подробно описаны в пунктах 2 и 3.

2. Подсистема синтеза речи по тексту

По функциональным признакам систему синтеза речи можно представить в виде четырёх процессоров (рис.3): лингвистического, фонетического, просодического и акустического.

Входными данными системы является орфографический текст, который последовательно обрабатывается каждым из процессоров, выходные данные - синтезированная речь.

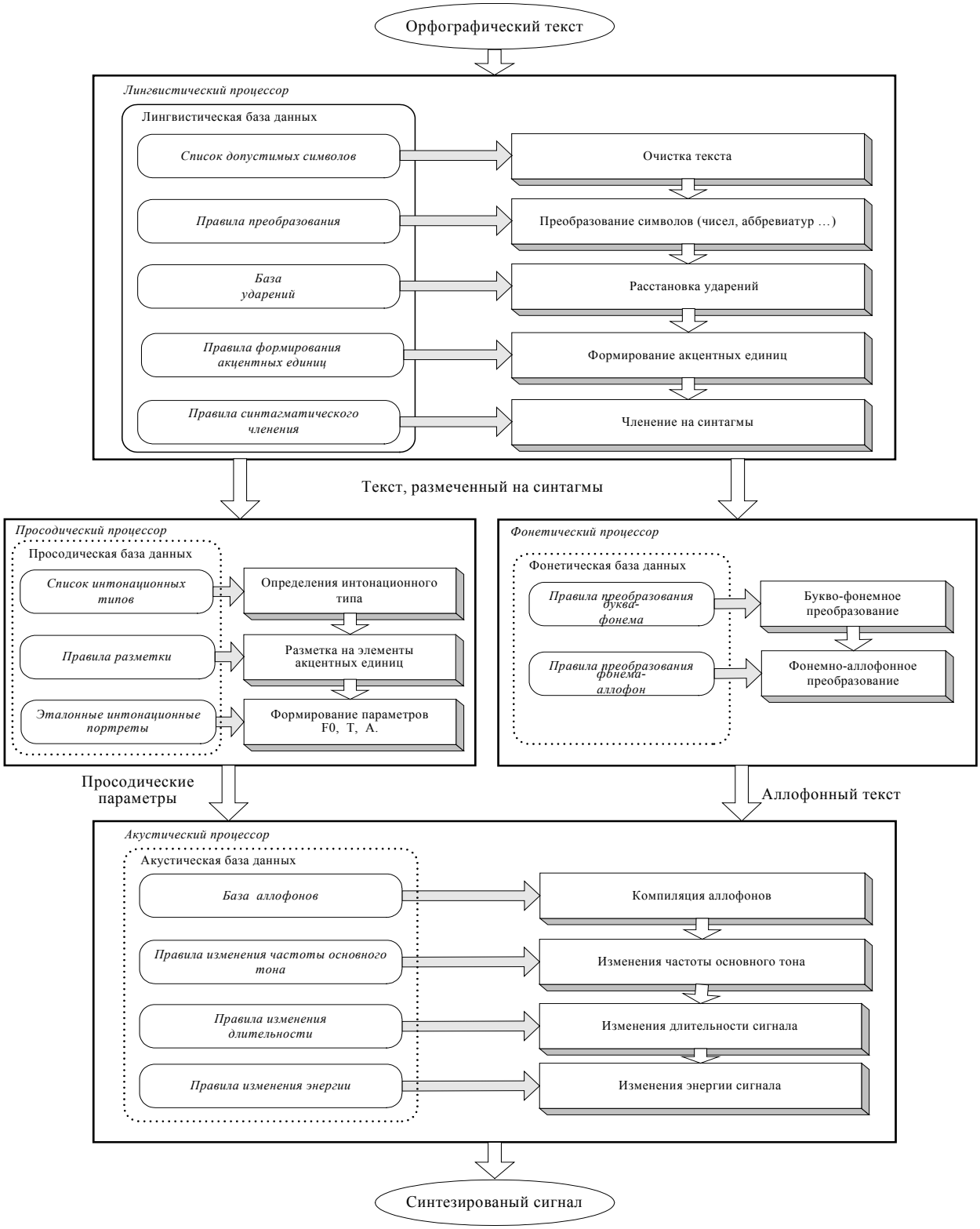

Рис. 3. Общая схема системы синтеза речи по тексту

Лингвистический процессор

Лингвистический процессор осуществляет следующую обработку: очистка текста, преобразование символов (аббревиатур, сокращений, чисел и др.), расстановка словесных ударений (на основе базы ударений и эвристических правил), формирование акцентных единиц (АЕ) и членение текста на синтагмы. Преобразованный текст, размеченный на синтагмы, поступает на вход просодического и фонетического процессоров.

Просодический процессор

Просодический процессор в первую очередь определяет интонационный тип каждой синтагмы, который зависит от её синтаксического статуса и грамматических категорий слов в составе синтагмы. В разработанной системе 14 различных интонационных типов: один тип восклицания, два типа вопроса, пять типов завершенности, шесть типов незавершенности. Кроме того, подтип синтагмы определяется количеством АЕ в ней.

Следующий этап просодической обработки – разбиение каждой АЕ на элементы АЕ (ЭАЕ): предъядро, ядро и заядро. Ядром АЕ является полноударный гласный, предъядром - все звуки, находящиеся слева от ядра, заядром - все звуки справа от ядра.

Затем происходит формирование просодических параметров каждого ЭАЕ синтагмы. Для этого используются эталонные интонационные «портреты» синтагм, хранящиеся в базе данных просодического процессора (просодической БД). Интонационный «портрет» для каждого типа синтагмы содержит контуры мелодики (F0), ритмики (Т) и энергии (А) в параметрическом виде. Полученные из просодической БД значения для каждого ЭАЕ приводятся к заданному частотному диапазону голоса.

Фонетический процессор

Фонетический процессор состоит из блоков преобразования буква-фонема и фонема-аллофон.

Первый из блоков использует правила преобразования буква-фонема для русского языка, в которых учитывается, в частности, позиция ударения в слове. При этом ударные гласные маркируются индексом 0, предударные - индексом 1, заударные - индексом 2. В служебных словах ударная гласная маркируется индексом 5, а порядок индексирования заударных и предударных, при их наличии, остаётся прежним.

Блок преобразования фонема-аллофон присваивает каждой фонеме, в зависимости от её типа и позиционных и комбинаторных особенностей расположения в слове, трёхзначный индекс (например, А043). Численное значение каждого индекса строго определено.

Акустический процессор

Акустический процессор, используя полученную от фонетического процессора информацию, компилирует отрезки соответствующих естественных звуковых волн, содержащихся в базе аллофонов. Затем, на основе полученной от просодического процессора информации, изменяет значения частоты основного тона, длительности и энергии полученного речевого сигнала.

3. Подсистема распознавания ключевых слов в потоке речи

Общая схема подсистемы распознавания речи и последовательность работы с речевым сигналом представлена на рис.4. Рассмотрим приведенную схему более подробно по блокам, входящим в нее.

Рис.4. Последовательность действий при обнаружении ключевого слова в потоке речи

Запись очередного блока

Этот элемент схемы введен для того, что бы подчеркнуть, что основной целью программы является поиск эталонов именно в реальном времени. Однако в программе имеется возможность поиска эталонов и в записи. При этом выполняются все те же действия, что и для реального времени, с одним отличием — нет разделения записи на блоки. Использование при записи блоков меньшего размера позволяет уменьшить задержку между реакцией программы и произнесенным эталоном. Запись ведется блоками по 40мс.

Вычисление сонограммы записанного сигнала

Вычисление сонограммы проводится путем пропускания исходной записи через набор полосовых фильтров Чебышева. Полосы пропускания фильтров выбраны в соответствии со шкалой в барках.

№ канала	1	2	3	4	5	6	7	8	9	10
$F_{\min}$, Гц	0	100	200	300	395	510	630	765	920	1075
$F_{\max}$, Гц	100	200	300	400	505	630	770	915	1080	1265
№ канала	11	12	13	14	15	16	17	18	19	20
$F_{\min}$, Гц	1265	1480	1710	1990	2310	2675	3125	3650	4350	5250
$F_{\max}$, Гц	1475	1720	1990	2310	2690	3125	3675	4350	5250	6350

Таблица 1. Полосы пропускания фильтров

Вычисление отфильтрованного сигнала выполняется по формуле:

$$y_i = \frac{B_0 x_i - \sum_{k=1}^{N} B_k x_{i-k} - A_k y_{i-k}}{A_0},$$

где x_i — входной, не отфильтрованный сигнал;

y_i — выходной, отфильтрованный сигнал;

A_k, B_k — заранее вычисленные коэффициенты фильтра

На рис.5 представлены амплитудно-частотные характеристики используемой гребёнки фильтров.

После вычисления отфильтрованного сигнала в каждой из указанных полос для оценки уровня сигнала вычисляется оценка его среднего квадратичного отклонения (СКО) по следующей формуле:

$$U = \sqrt{M(y^2) - (M(y))^2},$$

где $M(x)$ — математическое ожидание величины *x*.

Интервал, на котором будет вычисляться СКО, определяет пользователь, задавая параметр: частота спектральных срезов. Вычисленное значение умножается на соответствующий коэффициент полосы, задаваемый эквалайзером. Значения множителей вычисляются по формуле:

$$Mux = 10^{E_i + E_{All}/20}$$

где E_i — значение эквалайзера для i-го канала;

E_{All} — общее усиление.

Вычисленные значения усредняются на интервале, указанном пользователем.

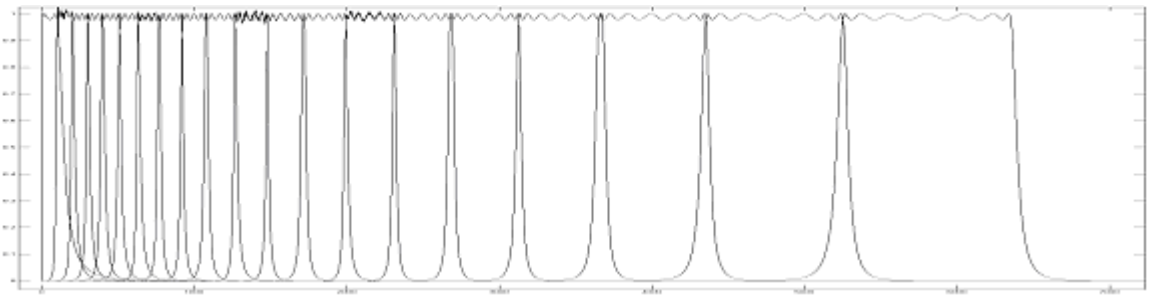

Рис. 5 Амплитудно-частотные характеристики используемой гребёнки фильтров

Нормирование сонограммы.

После вычисления сонограммы осуществляется её автоматическое нормирование. Нормированные значения вычисляются по формуле:

$$Sn(n,j) = \sum_{k=n-T}^{n+T} \sum_{l=0}^{20} \Delta(S(n,j), S(k,l)),$$

где $Sn(n,j)$ — нормированное значение точки n,j сонограммы;

$S(n,j)$ — ненормированное значение.

$$\Delta(S(n,j),S(k,l)) = \begin{cases} 1, & \text{если} \quad S(n,j) - S(k,l) > \varepsilon \\ 0, & \text{если} \quad -\varepsilon \le S(n,j) - S(k,l) \le \varepsilon \\ -1, & \text{если} \quad S(n,j) - S(k,l) < -\varepsilon \end{cases}$$

где ε — порог дифференцирования (по своему действию аналогичен общему усилению сигнала E_{All}).

НДП -вычисление интегральных расстояний

Разработанный алгоритм распознавания произносимых слов базируется на методе непрерывного ДП-сопоставления распознаваемого сигнала с эталонами (НДП-метод) [4]. Главным достоинством метода является то, что он позволяет определить вероятность присутствия слова в текущем речевом потоке и оценить его временное местоположение в реальных условиях наличия разного рода акустических помех. Метод дает возможность осуществить в процессе распознавания динамическое выравнивание временных шкал эталонного описания слова и его реализации в текущей речи при неизвестных начале и конце слова.

Рассмотрим процесс вычисления для одного эталона.

Пусть $\{\overline{E(m)}\} = \{E(0), E(1), \ldots, E(m), \ldots, E(M)\}$ есть последовательность векторов в эталоне слова, а $\{\overline{S(n)}\} = \{S(0), S(1), \ldots, S(n), \ldots, S(N)\}$ — последовательность векторов в текущем речевом потоке.

Первым шагом является нахождение матрицы локальных расстояний $d\{\overline{S(n)}; \overline{E(m)}\}$ между векторами эталона и текущего речевого потока:

$$d\{\overline{S(n)}; \overline{E(m)}\} = \frac{1}{E_0} \sum_{l=1}^{L} |S(n,l) - E(m,l)|,$$

где L — размерность векторов эталонного и текущего речевого потока (для нашего случая 20);

E_0 — делитель данного эталонного сигнала. Его нахождение будет рассмотрено позже.

Далее вычисляется матрица интегральных расстояний $D(n,m)$, матрица времен $T(n,m)$, и матрица переходов $Tr(n,m)$. Начальные условия для расчетов следующие:

$T(n,0) = 0$,

$T(0,m) = 0$,

$D(n,0) = d\{\overline{S(n)}; \overline{E(0)}\}$,

$D(0,m) = d\{\overline{S(0)}; \overline{E(m)}\} + D(0,m-1) + k * |m-1|$,

$Tr(n,0) = TrEnd$, для $n = \overline{0,N} \quad m = \overline{1,M}$

Новые значения $D(n,m)$, $T(n,m)$ и $Tr(n,m)$ вычисляются в соответствие с формулами:

$$D(n,m) = \min \begin{bmatrix} D(n-1,m) + k_h d\{\overline{S(n)}; \overline{E(m)}\} + \frac{k}{M}|m - T(n-1,m)|; & (1) \\ D(n,m-1) + k_v d\{\overline{S(n)}; \overline{E(m)}\} + \frac{k}{M}|m-1 - T(n,m-1)|; & (2) \\ D(n-1,m-1) + k_d d\{\overline{S(n)}; \overline{E(m)}\} + \frac{k}{M}|m-1 - T(n-1,m-1)|; & (3) \end{bmatrix},$$

$$T(n,m) = \begin{cases} T(n-1,m)+1, & \text{если} \quad D(n,m) = (1) \\ T(n,m-1), & \text{если} \quad D(n,m) = (2) \\ T(n-1,m-1)+1, & \text{если} \quad D(n,m) = (3) \end{cases},$$

$$Tr(n,m) = \begin{cases} TrHoriz, & \text{если} \quad D(n,m) = (1) \\ TrVert, & \text{если} \quad D(n,m) = (2) \\ TrDiag, & \text{если} \quad D(n,m) = (3) \end{cases},$$

где k — вес времени, задаваемый пользователем, нормированный к длине эталона,

$Tr(n,m)$ — матрица переходов, носящая вспомогательную для распознавания функцию, состоящую в потенциальной возможности нахождения временного соответствия эталонного сигнала и произнесенного (перенос меток).

В связи с тем, что для длинных эталонов интегральное расстояние будет, как правило, больше чем для коротких, в вычисления локальных расстояний вводится описанный ранее делитель E_0, предназначенный для нормирования графиков длинных эталонов и коротких эталонов

Поиск локальных минимумов НДП-графиков

После вычисления НДП-графики интегрального расстояния (рис. 6) анализируются на наличие локальных минимумов — точек, имеющих минимальное значение в некоторой окрестности. Окрестность составляет 10 точек и может быть изменена перекомпиляцией программы. Для того, что бы найденный минимум принимался к рассмотрению, он должен иметь меньшее значение, чем заданный пользователем порог (порог нахождения минимумов).

Рис. 6. Анализ графиков интегрального расстояния

Поиск среди минимумов кандидата- лидера

Среди найденных минимумов на некотором промежутке ищется такой, который бы имел наименьшее из всех значение и был меньше, чем порог распознавания, задаваемый пользователем.

Обработка случаев двух близких минимумов

Предварительно, до запуска программы, проводится анализ поведения графиков интегральных расстояний для случаев произношения различных эталонов. Далее, при нахождении минимума-лидера, проверяется, к какому эталону он относится, и анализируются минимумы-конкуренты. В некоторых случаях минимум-конкурент может занять позицию лидера. После данного этапа эталон считается распознанным.

Заключение

Описанная демо-версия системы является гибкой и легко настраиваемой. Система позволяет добавлять речевые ресурсы: эталоны ключевых слов для распознавания и базы аллофонов для синтеза речи, и модифицировать текстовые ресурсы: изменять ключевые слова, добавлять сценарии диалогов.

Тестирование демо-версии показало, что точность распознавания ключевых слов в потоке речи достигает 95% при предварительной настройке на диктора. Синтезированная речь приближается по звучанию к естественной.

Дальнейшее совершенствование системы будет направлено на повышение точности дикторонезависимого распознавания речи.

Литература

1) Лобанов Б.М. и др. Синтезатор персонализированной речи по тексту “ЛобаноФон-2000” // Труды Международной конференции, посвящённой 100-летию российской экспериментальной фонетики.- Ст.-Петербург, 1 – 4 февраля 2001 г.С. 101-104.
2) Lobanov B.M., Tsirulnik L.I. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS // Proc.of the Ninth International Conference “Speech and Computer” SPECOM’2004, Saint-Petersburg, 2004, p. 17-21.
3) Лобанов Б.М., Давыдов А.Г., Киселёв В.В., Цирульник Л.И. Система сегментации речевого сигнала методом анализа через синтез Известия Белорусской инженерной академии №1(17)/1’2004, Минск, С.112 – 115.
4) Lobanov B.M., Levkovskaya T.V. Continuous Speech Recognizer for Aircraft Application // Proceedings of the 2nd International Workshop “Speech and Computer” – SPECOM’97.-Cluj-Napoca, 1997.-P. 97-102.