УДК 621.391

СИСТЕМА СИНТЕЗА РУССКОЙ РЕЧИ НА ОСНОВЕ КОМПИЛЯЦИОННОГО МЕТОДА

В.В. КИСЕЛЕВ, Б.М. ЛОБАНОВ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Объединенный институт проблем информатики НАН Беларуси
Сурганова, 6, Минск, 220013, Беларусь

Поступила в редакцию 3 августа 2004

Описывается разработанная система синтеза русской речи на основе компиляционного метода. Освещена проблема выбора минимального акустического сегмента в компиляционном типе моделирования речевого сигнала. Дается понятие аллофона, поскольку именно аллофон взят за основу в акустическом инвентаре системы. Представлено пошаговое описание лингвистической обработки текста, включающее в себя очистку текста, различные преобразования (числительные, аббревиатуры, сокращения и т.д.), расстановку ударений, формирование акцентных групп, членение на синтагмы. Описаны просодические характеристики русской речи, приведены классификации синтагм и типы акцентных групп. Показан основной принцип построения просодического и акустического модуля в разработанной системе синтеза русской речи.

Ключевые слова: речь, синтез речи, компиляционный метод, аллофон, дифон.

Введение

Речь, пожалуй, является одним из древнейших средств коммуникации между людьми и принадлежит к высшей, интеллектуальной деятельности человека. Стремительный рост вычислительной техники и высокотехнологический прорыв в середине 80-х годов XX в. (появления теории цифровой обработки сигналов, методов анализа естественных языков) позволил сформировать огромный научный потенциал для создания высококачественных систем синтезирования речи посредством выделения отдельного метода — компиляционного, главная идея которого заключается в соединении готовых, заранее подготовленных, минимальных речевых единиц.

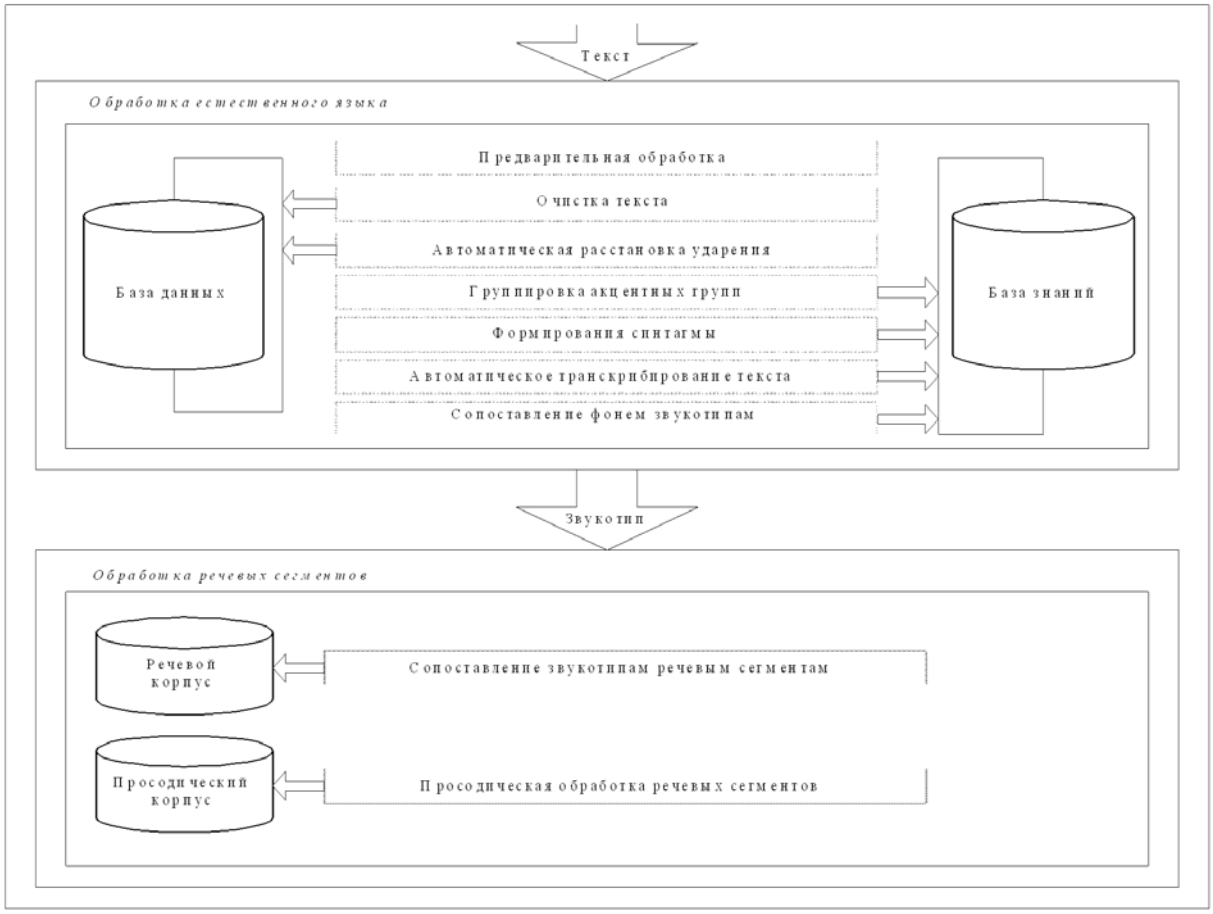
С общей точки зрения такой метод является наиболее простым решением для достижения натуральности и естественности звучания, однако при реализации и непосредственном конструировании подобного рода систем возникает ряд трудностей. Выбор минимальной единицы, пожалуй, является одной их наиболее труднопреодолимой преградой.

Обычно выбор является компромиссным вариантом между более длинными и более короткими сегментами. Естественно, что чем больше длинна сегмента (например, слово), чем меньше точек соединения, тем более естественней и натуральней звучит синтетическая речь, однако в этом случае словарь возможных сообщений сводится к возможному семантически осмысленному сочетанию данных сегментов, следовательно, о неограниченном словаре говорить не приходится.

С коротким сегментом требуется значительно меньше памяти, однако затруднен этап сборки синтезированного сигнала и его обработки. По разным источникам выделяют следующие основные минимальные единицы: аллофоны, дифоны, трифоны, полуслоги, слоги и их различные интерпретации [1]. В разработанной системе синтеза русской речи в качестве минимальной акустической единицы используется смешанный тип сегментов, а именно аллофоны и дифоны. Рассмотрим более подробно понятие аллофон.

Каждая фонема в реальной речи подвергается различным модификациям в силу коартикуляции, а именно, вследствие комбинаторных звуковых изменений типа аккомодации и ассимиляции, связанных с ближайшим окружением фонемы, а также позиционных звуковых изменений типа редукции, обусловленных, прежде всего ее реализацией в ударном или неударном слоге. Таким образом, возникают фонетически обусловленные комбинаторные и позиционные варианты данной фонемы, называемые аллофонами. Понятие дифона не столь отчетливо. По разным источникам дифоном называют соединение двух аллофонов или же представление лишь стыков аллофонов, с разрывом в середине аллофона. Таким образом, в первом случае, при конкатенации двух дифонов процедура аналогична конкатенации аллофонов, во втором же случае механизм склейки более сложен. Соответственно модель, основанная на трифонах строится на таких же принципах.

Систему синтеза можно разбить на два больших модуля, это модуль обработки естественного языка и модуль обработки цифрового сигнала. Общая схема системы приведена на рисунке.



Общая схема системы синтеза речи

Модуль обработки естественного языка работает с входным орфографическим текстом, занимается его анализом и обработкой. Непосредственно в прикладных системах модуль естественного языка делится на три подмодуля: лингвистической, фонетической и просодической

обработки, в свою очередь модуль обработки цифрового сигнала заключается лишь в акустическом подмодуле.

В лингвистический модуль включены такие подмодули, как очистка текста, расшифровка числительных, даты и времени, аббревиатур, сокращений, Интернет ресурсов, автоматическая расстановка ударений, объединение слов в акцентные группы и членение на синтагмы [2]. Распишем более подробно эти этапы.

1. На первом шаге входной орфографический текст подвергается очистке от символов, не принадлежащих словарю. Словарь находится в базе данных системы и представляет собой список русских и английских букв, цифры и список специальных символов. Анализ текста происходит посимвольно, если *i*-й символ не принадлежит списку допустимых символов, то алгоритм переходит на (*i*+1)-й символ.

2. Затем образуются орфографические синтагмы. Орфографической синтагмой считается последовательность слов, объединенных по правилам:

от начала текста до первого знака препинания;

от знака препинания до знака препинания;

от знака препинания до конца текста;

Список знаков препинания {‘,’ ‘.’ ‘:’ ‘;’ ‘-‘ ‘?’ ‘!’}.

Расшифровка сокращений, аббревиатур и числительных по определенным правилам.

Автоматическая расстановка ударения. В реализованной системе ударения расставляются посредством поиска необходимого слова в базе ударения. Если в базе присутствует слово, и оно не принадлежит к списку служебных слов, а именно союзам, предлогам и частицам, тогда ему присваивается полное ударение (маркировка индексом 0). Если слово принадлежит списку служебных слов, тогда ему присваивается частичное ударение (маркировка индексом 5). Если слово не нашлось, тогда все гласные маркируются как частично ударные (индекс 5).

Объединение слов в акцентные группы происходит посредством объединения слов с полным ударением со словами, имеющими частичное ударение (служебные слова). Если слово принадлежит списку предлогов или списку союзов, тогда это слово присоединяется к последующему слову. Если слово принадлежит списку частиц (кроме частицы не), то оно присоединяется к предшествующему слову.

Формирование микро-синтагм. Формирование микро-синтагмы зависит от количества полных ударений в орфографической синтагме. Если количество полных ударений меньше или равно четырем, орфографическая синтагма остается без изменений. Если количество полных ударений больше четырех, необходимо разбить синтагму на микро-синтагмы. Первым информативным маркером для членения являются союзы и, или. Если данные союзы присутствуют в синтагме, тогда перед ними синтагма разделяется. После этого определяется количество полных ударений в каждой из синтагм. Если в какой-либо синтагме количество полных ударений превышает четырех, необходимо расчленить синтагму. Для этого вначале расставляются маркеры, где слова не могут быть разделены. К классу таких слов относятся прилагательные. Тип слова определяется через базу ударения. Если в базе ударения отсутствует информация о типе слова, тогда прилагательное определяется с большой степенью вероятности по окончанию (*ая, ее, его, ей, ему, ею, ие, ий, ими, их, ою, ого, ое, ой, ому, ою, ую, ый, ые, ым, ыми, ых, юю, яя*). После того как расставлены маркеры неделимых слов, происходит формирование микро-синтагм. Конец первой синтагмы ставится после третьего слова с полным ударением, если за ним не стоит маркер неделимости. В случае, если маркер стоит, конец синтагмы ставится после второго слова, и так же проверяется на наличие после него маркера неделимости. Данный цикл происходит итерационно до тех пор, пока в синтагме не останется меньше четырех полных ударений.

Фонетическая обработка заключается в автоматическом транскрибировании текста в фонемный вид, а затем фонемного текста в аллофонный. 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 остается прежний. Более детально процесс преобразования в аллофонный вид расписан в [3].

Просодическое оформление текстовой информации, в лингвистической обработке, заключается в сопоставлении интонационных контуров к соответствующим типам синтагм. Для этого необходимо классифицировать акцентные группы и разделить их на составляющие.

Акцентные группы ранжируются относительно границ синтагмы на конечные, начальные и серединные. Эти категории акцентных групп вносят различный по значимости вклад в формирование просодического контура синтагмы. Основное разнообразие мелодических контуров реализуется на конечной акцентной группе, несущей синтагматическое ударение, существенно меньшее разнообразие на начальной и срединной. Мелодические, ритмические и энергетические характеристики акцентных групп являются теми минимальными единицами, из которых складывается интонация синтагмы, фразы и текста в целом.

Коротко рассмотрим далее принципы мелодического, ритмического и энергетического оформления речи для различных интонационных типов.

Как уже указывалось, акцентная группа является той минимальной единицей, на которой задаются интонационные характеристики и из совокупности которых складывается интонационный тип синтагмы. Практически совершенно необходимым для синтеза речи по тексту является изучение тех интонационных типов, которые связаны с грамматической (синтаксической) функцией интонации. Среди них рассмотрим вопросы реализации наиболее важных интонационных типов, таких, как интонация завершенности, незавершенности, вводности, перечисления, вопроса и восклицания.

Проведенные автором исследования показали, что для вполне удовлетворительного задания указанных типов интонации синтагмы достаточно выделить три класса акцентных групп: конечные, серединные и начальные. Интонация синтагмы складывается последовательным соединением указанных классов акцентных групп, на каждом из которых в соответствии с требуемым интонационным типом формируются необходимые контуры мелодики, ритмики и энергетики. При наличии в синтагме более трех акцентных групп добавляется необходимое количество серединных групп, а при наличии менее трех групп исключается вначале серединная, а затем начальная акцентная группа.

Мелодика, ритмика и энергетика акцентной группы задаются нормированными значениями частоты, длительности и интенсивности на трех ее участках: ядре, предъядре и заядре. Ядром акцентной группы является ударный слог, отмеченный знаком группового ударения, Предъядром и заядром — соответственно предшествующие ему и следующие за ним фонемы акцентной группы.

Главная задача **акустического модуля** — генерация (синтез) речевого сигнала на основе трех типов параметров[4]:

просодических параметров (F_0 — частота основного тона, T — длительность звуков, A — амплитуда звуков), которые поступают от просодического подмодуля;

фонетических параметров, поступающих от фонетического подмодуля (в зависимости от типа фонетического процессора эти параметры могут быть различными: формантными параметрами (F1, F2, A1 …), параметры сечений речевого тракта, номер аллофона или сегмента и т.д.);

параметры синтезируемого голоса, обеспечивающие желаемую тембровую индивидуальность.

Разработанная система найдет свое применение в интерактивном взаимодействии пользователя и ЭВМ, поможет незрячим людям, а также растущий интерес к компьютерным играм, развлекающим программам показывает потребность и применимость синтезатора речи.

TTS SYSTEM FOR RUSSIAN SPEECH BY CONCATENATION METHOD

V.V. KISELOV, B.M. LOBANOV

Abstract

The developed TTS system of Russian speech on a basis concatenation a method is described in this paper. The problem of a choice of the minimal acoustic segment in concatenation type of modeling of a speech signal is covered. The concept allophone as allophone it is taken for a basis in acoustic inventory is given. The step-by-step describe of linguistic processing text, including clearing of the text, various transformations (numerals, abbreviations), arrangement of accents, formations of accent groups, partitioning on syntagmas are submitted. It is described prosody characteristics of Russian speech; classifications of syntagmas, and types of accent groups are resulted.

Литература

1. *Kishore S.P., Black A.W.* // EUROSPEECH 2003. P. 1317–1320.
2. *Киселев В.В., Лобанов Б.М., Левковская Т.В., Хейдоров И.Э.* // Тр. междунар. конф., посвященной 100-летию российской экспериментальной фонетики. СПб., 2001, С. 101–104.
3. *Киселев В.В., Шаков А.В.* // Сб. тр. "Автоматическое распознавание и синтез речи". Мн., ИТК НАН Беларуси, 2000. С. 155–163
4. *Бухтилов Л.Д., Лобанов Б.М., Минкевич В.В.* // Автоматическое распознавание слуховых образов: АРСО-10. Тбилиси, 1978. С. 132–133.