СОВРЕМЕННЫЕ ТЕХНОЛОГИИ СИНТЕЗА УСТНОЙ РЕЧИ

О.А. Русанова

Красноярский государственный технический университет,
кафедра Вычислительной техники

rusanova@fivt.krasn.ru

Введение

Вопросы синтеза и распознавания речи человека компьютером становятся все более актуальными. Речевые технологии уже внедряются в нашу жизнь. Успехи в развитии естественно-языковых технологий обещают широкий доступ к on-line информации и электронным сервисам. Так как почти каждый говорит и понимает речь, развитие естественно-языковых систем позволит человеку без специальных навыков общаться с компьютером в любое время и в любом месте без дополнительного обучения, используя такие устройства, как мобильный телефон, получать доступ к информации или к управлению устройствами. Сотовый телефон так же можно использовать как устройство перевода речи в речь, которое распознает сказанную вами фразу на одном из трех языков – японском, английском и немецком, и переводит на любой их этих языков.

В настоящее время вопросом синтеза речи занимается большое число исследовательских групп, каждая из которых создает свое описание речевого сигнала, и в конечном итоге - программный продукт: Клуб голосовых технологий МГУ и фирма ПРОМТ - “Magic Goody” [19], Sakrament, г. Минск [20], Microsoft Speech SDK [16], AT&T [17], Verbmobil Германского исследовательского центра искусственного интеллекта [9]. На данный момент из программных пакетов, поддерживающих русский язык, наиболее широко распространены Microsoft Speech SDK, Lernout&Hauspie и разработка “Digalo” фирмы Elan Informatique [10] (см. табл. 1).

Несмотря на широкий набор разработок, проблема синтеза речи до сих пор считается решенной лишь удовлетворительно. До сих пор нет общего мнения, какой из существующих подходов дает наилучшие результаты, какие модели синтеза речи являются наиболее перспективными. В этой статье проведен обзор систем синтеза речевых сообщений и проведена классификация решений вопросов синтеза речи.

1. Компании, занимающиеся вопросами синтеза речи

Для того, чтобы разработки разных групп, занимающихся вопросами синтеза речи, могли быть совместимы между собой, был создан единый стандарт Speech API (Application Programming Interface) – интерфейс создания речевых приложений. Не удивительно, что он был предложен фирмой Microsoft – лидером в области разработки программного обеспечения.

Критерием оценки успешности той или иной разработки может служить ее спрос на рынке речевых технологий, ее коммерческая реализуемость. Фирма Microsoft приводит свой список компаний, занимающихся производством программного обеспечения на данном сегменте рынка [16]. Естественно, что это те производители, которые поддерживают стандарт SAPI. Но для полной картины современного состояния вопроса необходимо принять во внимание и те разработки, которые являются закрытыми для других разработчиков.

В основе любой речевой технологии лежит так называемый «engine» или ядро программы синтеза речи – набор данных и правил, по которым осуществляется обработка данных. В зависимости от назначения этого ядра различают TTS и ASR engine. TTS (Text-to-Speech) engine предоставляет возможность синтеза речи по тексту, а ASR (Automatic Speech Recognition) engine – распознавания речи. Некоторые производители ставят своей задачей разработку только ядра программы синтеза речи, другие – разработку ядра и решений на их основе. Некоторые компании используют чужие engines для разработки своих приложений. В табл. 1 приведены разработчики, уже создавшие свои системы синтеза речи и решений на их основе. Таблица дает общее представление о наполненности рынка речевых технологий, в ней указаны фирма производитель, ее разработка, назначение продукта, цена, объем ядра, поддерживаемая версия SAPI, языки и дана краткая дополнительная информация. Список фирм производителей дает представление о географии разработок, назначение продукта – о потребности в речевых технологиях, цена – о качестве разработки и об ожидаемом коммерческом эффекте от ее применения, объем ядра – о возможных сферах его применения, версия SAPI – об уровне компании-разработчика, а поддерживаемые языки – о географии интереса потребителей.

Наиболее популярным приложением речевых технологий на сегодняшний момент является создание: серверных приложений, решения для персональных компьютеров и создание встраиваемых программ. Серверные приложения – это в первую очередь создание call-центров, организация сетевой телефонии (IP-телефонии). Самый широкий спектр применений у решений для персональных компьютеров: озвучивание игр, чтение почты или другой текстовой информации. Компания Conversay предоставляет возможность организации получения информации из Интернета голосовыми командами [2]. Elan Informatique специализируется на разработке приложений для слепых и слабовидящих людей. Но все большую популярность получает разработка «легких» ядер – систем, обладающих низкими системными требованиями и позволяющих создавать встраиваемые приложения. Такой продукт можно встретить и в мобильном телефоне, и в palm-top’е, и в системе управления автомобилем. Уже создана стиральная машина, рассказывающая хозяину, что и как нужно сделать, что бы постирать белье, загруженное в ее барабан.

Несмотря на растущую популярность создания речевых приложений, из табл. 1 видно, что производителей, заинтересованных в разработке ядер для синтеза русской речи и уже создавших их, немного. Это Microsoft, Elan Informatique, L&H, Sakrament и Клуб голосовых технологий при Научном парке МГУ.

Помимо компаний, уже создавших свои первые версии средств создания речевых приложений, существуют и те, кто таких результатов не успел добиться, но активно работает в этом направлении, либо компании, работающие в смежных областях, разработки которых могут быть полезны при решении вопросов синтеза речи. Это Бийский технологический институт, совместно с Томским университетом систем радиоуправления и радиоэлектроники; объединение «Интеллект» МГУ им. М.В.Ломоносова; «Центр речевых технологий», г. С-Петербург; «Истра-софт», г. Истра; компания Etaco Inc., Нью-Йорк и другие.

Бийским Технологическим Институтом Алтайского государственного технического университета совместно с кафедрой Компьютерных систем в управлении и проектировании Томского университета систем управления и радиоэлектроники проводится работа по созданию систем синтеза речи. За это время ими теоретически разработаны некоторые подходы по структуризации и формализации как речевого сигнала, так и орфографического текста, алгоритмы преобразования текста в речевой сигнал по правилам с использованием параметров речеобразующего тракта человека. На практике существуют автоматическая расстановка ударений для русского языка (макет), фонетическое транскрибирование для русского и английского языка (рабочий вариант). При этом производится расчет просодических характеристик речевого сигнала: расчет длительностей звуков и частоты основного тона (рабочий вариант). В целом предполагается создание аппаратно-программного комплекса по преобразованию печатного текста в речь [4].

Объединение «Интеллект» кафедры Математической теории интеллектуальных систем и лаборатории Проблем теоретической кибернетики Механико-математического факультета МГУ им. М.В.Ломоносова проводит теоретические и практические исследования по следующим основным направлениям: дискретные алгоритмы распознавания речи; устойчивое к шумам и помехам распознавание речи с использованием дополнительных (неакустических) источников речевой информации; грамматики естественных языков и их применение в системах распознавания речи; автоматическое чтение по губам; синтез речи [5].

«Центр речевых технологий», г. С-Петербург – это небольшая группа инженеров-разработчиков, работавших в крупнейших Научно-исследовательских институтах военно-промышленного комплекса Санкт-Петербурга. Основное поле деятельности данной группы – это распознавание речи; верификация голоса (разграничение доступа с использованием голоса); шумоочистка речевых сигналов; судебные фоноэкспертизы и расшифровка "черных ящиков"; сжатие речи [22].

Разработки "Истра-Софт" в области речевых технологий включают в себя следующие основные направления: сжатие речевых файлов, распознавание речи, синтез речи по тексту, идентификация личности по голосу. На данный момент компания успела выпустить такие программные продукты как серию "Профессор Хиггинс": "Профессор Хиггинс. Английский без акцента!" и "Профессор Хиггинс. Русская фонетика", программу для детей "Остров арифметики". Помимо этого выпущено программное обеспечение для распознавания фо-

нем, Интернет-телефонии, передачи речи по корпоративным сетям, голосовых писем. «Истра-Софт» разрабатывает системы обработки речи (сжатие, распознавание, синтез). Компанией разработана многопотоковая программа сжатия речи, основанная на выделении и распознавании фонем. Программа предназначена для встраивания в программные системы передачи речи по Интернету и Интранету, для локальной сети предприятия и цифровой телефонии, для любой системы цифровой передачи данных. На основе этих разработок создан модуль диктофона, позволяющий сжать звуковое письмо. Программа дает возможность записать три часа речи на 3-х дюймовой дискете. Модуль предназначен для встраивания в почтовые программы для записи и воспроизведения звуковых писем и передачи их по Интернету и любым другим каналам передачи цифровых данных. В настоящее время ведется разработка программы голосонезависимого командного распознавания. Проблемы синтеза речи по тексту «Истра-Софт» планирует решать на следующем этапе разработок, так как для каждого языка пока необходимо набрать библиотеку описаний фонем и базы данных соответствия транскрипции и буквенного описания. Компания планирует реализацию алгоритма сжатия и восстановления звуковых файлов на базе своей микросхемы [5].

Компания Ectaco Inc. с 1990 г занимается лингвистическими исследованиями и разработкой карманных электронных переводчиков. Инженеры, лингвисты и программисты Ectaco Inc. выполняли проекты для таких компаний, как Franklin Electronic Publishers, L&H, C-Pen, G-Data, Adobe и др. Одна из последних технологий, разработанная в Ectaco Inc., основана на параметрическом (формантном) описании, которое моделирует звуковой тракт человека. Этот принцип дает системе, в первую очередь, компактность, то есть возможность ее использования в современных карманных компьютерах. Также появляется возможность моделирования мужских, женских и детских голосов с разными тембрами, темпом, тональностью произнесения и пр. Форматная модель позволяет и более аккуратно обрабатывать голос для передачи по каналам связи с сильными искажениями (например, по телефонным) [11].

2. Классификация систем синтеза речи

Классификацию подходов к синтезу речи можно провести по нескольким признакам: по характеру синтезируемой речи, по принципу построения синтезируемых сообщений, по методу синтеза и по принципу реализации (рис. 1).

Синтезаторы речи различают прежде всего по *характеру синтезируемой речи*. Это может быть или предварительно закодированная, сжатая по возможности речь, или искусственные речеподобные звуки, сформированные электронным устройством.

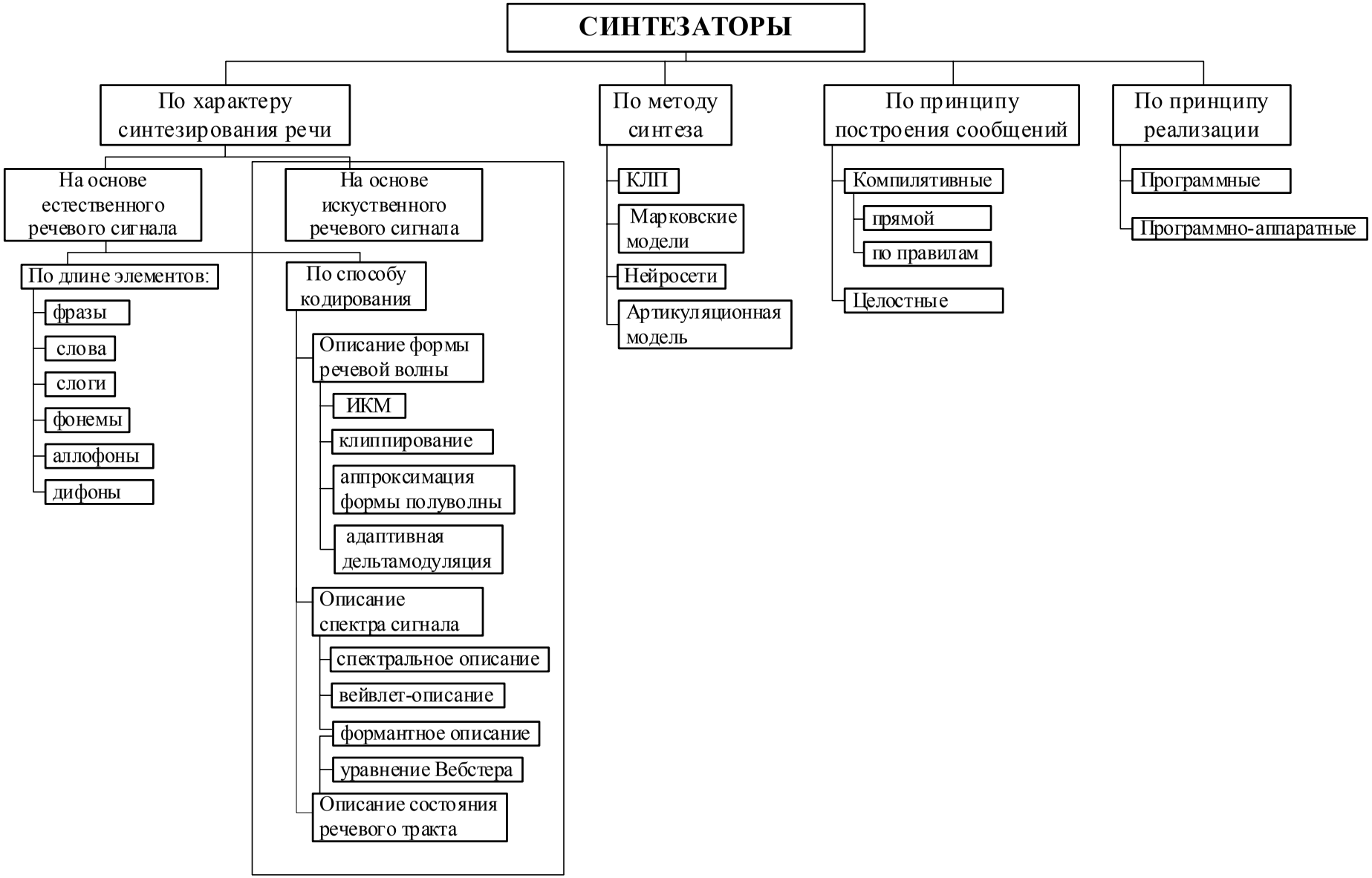

Рис. 1. Классификация синтезаторов речевых сообщений

Синтезаторы на основе естественной речи используют заранее записанные речевые сообщения, которые могут делиться на фразы, слова, слоги, фонемы, дифоны и аллофоны. В этом случае для уменьшения объема занимаемой памяти применяют широко известные из общей теории обработки сигналов способы сжатия сигнала: описание формы речевой волны, спектральное описание и описание состояния речевого тракта.

Для сжатия сигнала по описанию формы речевого волны используют: логарифмическую импульсно-кодовую модуляцию (ИКМ), клиппирование, аппроксимацию формы полуволны, адаптивную дельта-модуляцию и всевозможные комбинированные способы.

Но самое сильное сжатие описания любого процесса — выделение факторов (параметров), порождающих данный процесс. С этой точки зрения предпочтение надо отдать описанию спектра сигнала и описанию состояния речевого тракта. Речевой тракт описывается уравнением Вебстера, а спектр вычисляется при помощи разложения Фурье или вейвлет-преобразованию. Формантное описание – тоже одно из самых эффективных, лежит на стыке решения уравнения Вебстера и спектрального описания (используется компанией Etaco [11]).

Необходимо отметить, что речевой сигнал, синтезированный по одному из вышеперечисленных методов, так же можно считать искусственным.

Еще одним критерием классификации синтезаторов служит *метод синтеза речевого сигнала.* Среди методов можно выделить использование Марковских моделей, нейросетей, применение артикуляционной модели, кодирования с линейным предсказанием (КЛП-синтез) и эвристических правил. Как правило, использование Марковских моделей и нейросетей оправдано для систем, направленных как на синтез, так и на распознавание речи. Часто разработчики комбинируют несколько подходов для решения этих задач, как, например, в системе Verbmobil [9].

По принципу построения сообщений выделяют компилятивные и целостные синтезаторы.

Целостные синтезаторы реализуют прямой синтез речи - сначала естественная речевая информация записывается в виде единичных слов или фраз, а затем при выводе заданная речь воссоздается путем соответствующего комбинирования ранее записанных речевых единиц. Для метода прямого синтеза характерно оперирование большими объемами данных, значительно превышающими объемы, используемые в вышеописанных методах, но при этом он обеспечивает весьма высокое качество речи. Этот метод синтеза применяется в системах автоматического голосового оповещения и в других областях с небольшим количеством используемых слов (порядка нескольких десятков).

Компилятивные синтезаторы создают произвольные речевые сообщения по тексту из отдельных элементов естественного или искусственного происхождения. Компилятивный синтез может быть прямым или по правилам.

Прямой компилятивный синтез уступает другим видам синтеза по натуральности речи, так как простая стыковка звуков не соответствует физике речеобразования и плохо воспринимается на слух, а для учета важнейших законо-

мерностей речеобразования пока еще нет достаточно простых и эффективных моделей.

Для улучшения качества звучания используется компилятивный синтез по правилам. В общем случае система должна иметь достаточно полный набор исходных речевых элементов, а также иерархическую систему правил преобразования и объединения элементов в более крупные отрезки речи на уровнях звуков, слов и фраз с учетом просодических явлений (интонации, паузы, словесные и фразовые ударения), иметь возможность накопления и использования опыта разработчиков-лингвистов, сконцентрированного в виде баз знаний.

Проще всего осуществлять такой синтез по фонетической транскрипции, в противном случае система должна обладать способностью преобразовывать орфографический текст в фонетический, а соотношение между ними очень сложное. Для такого преобразования текст сначала подвергается лингвистической обработке, заключающейся в синтаксическом анализе и различении отдельных слов, а затем определяются способы чтения отдельных слов. После этого слова, являющиеся объектом синтеза, разделяются на единицы компиляции; к ним добавляется информация о качественных характеристиках звука и производится синтез. Использование набора правил дает возможность формирования естественного просодического оформления высказываний [24]. В системах компилятивного синтеза в качестве элементов компиляции используются различные типы единиц: аллофоны, дифоны, слоги, полуслоги, двуслоги и т.д. Основная проблема при выборе единиц синтеза — это учет коартикуляции, т. е. взаимного влияния артикуляционных движений при произнесении соседних звуков и, как следствие, зависимости параметров фонемы от фонетического окружения. Особенно подвержены коартикуляции переходные (начальный и конечный) участки фонем. Фактор коартикуляции вынуждает использовать аллофоны — разновидности фонем, обусловленные конкретным звуковым контекстом. Примером такого подхода может служить разработка филологического факультета МГУ[24].

Системы компилятивного типа пока уступают целостным системам в натуральности звучания. Однако они имеют очень важные преимущества: компактность описания сообщений, неограниченный словарь, возможность синтеза речи по тексту.

Еще одним критерием классификации синтезаторов является *принцип реализации,* который может быть программным и программно-аппаратным, не зависимо от математического аппарата, используемого при построении системы синтеза. Все примеры, приведенные выше можно отнести к программной реализации, а примером программно-аппаратной реализации может служить разработка фирм Dec и Fonix [14]

Библиографический список

1. http://actor.loquendo.com/actordemo/default.asp
2. http://conversay.com/products/default.aspCompany/default.asp
3. http://cslu.cse.ogi.edu/HLTsurvey/ch5node4.html

4. http://iclub.kemsu.ru/ts.
5. http://intsys.msu.ru/matis/lab.
6. http://isabase.philol.msu.ru/SpeechGroup/.
7. http://www.alantts.com/accueil.html.
8. http://www.blind.ru .
9. http://www.dfki.de/~wahlster.
10. http://www.digalo.com.
11. http://www.ectaco.com/.
12. http://www.flexvoice.com/demo.html.
13. http://www.fonix.com/downloads/ttsdemo.php.
14. http://www.fonix.com/products/dectalk/products.asp.
15. http://www.istrasoft.ru.
16. http://www.microsoft.com/speech/.
17. http://www.naturalvoices.att.com/demos/.
18. http://www.pholol.msu.ru/~oupl/new/main/index.php.
19. http://www.promt.ru.
20. http://www.sakrament.com/products/tts/.
21. http://www.scansoft.com/realspeak/.
22. http://www.speechpro.ru/.
23. http://www.speechworks.com/products/tts/speechify_demos/voicesof.cfm.
24. Зиновьева Н.В., Кривнова О.Ф., Захаров Л.М. Программный синтез русской речи (синтезатор «АГАФОН»)// Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог'95». Казань, 1995.