УДК 5019.7

МЕТОДЫ АНАЛИЗА И ПРОЕКТИРОВАНИЯ СИСТЕМЫ СИНТЕЗА ИСКУССТВЕННОЙ РЕЧИ

© Амиргалиев Е.Н., Мусабаев Р.Р.

КазНТУ им. К.И. Сатпаева, Казахстан, Алматы

e-mail: : *amir_ed@mail.ru, rmusab@gmail.com* ,

Abstract. In the given work the practical experience of speech synthesis algorithms realization and application are examined. Human-machinery speech interfaces, designated for the linguistics courseware is the subject of one of the most effective application of the given algorithms.

Введение

В данной работе рассматриваются вопросы создания информационной системы анализа и синтеза искусственной речи, некоторые проблемы которой приведены в предыдущих статьях [1, ?]. Главная задача систем синтеза речи – это преобразование входной текстовой информации в её выходное звуковое представление на одном из естественных языков. В самом оптимальном варианте, данные системы призваны заменить человека в процессе чтения текстов и выступить в роли автоматизированных дикторов, которые готовы неустанно читать тексты любых объёмов и любой сложности на любых языках. Под естественной речью понимаются звуковые и мимические составляющие речи человека. Входная текстовая информация обычно представляется в виде электронного текста (Рис. 1).

Данные алгоритмы в основном применяются при построении систем с человеко-машинными интерфейсами. Особенно их применение оправдано, если система оперирует текстовой информацией, подлежащей озвучиванию, а данная информация является переменной во времени. Классическим примером систем данного типа выступают системы контроля и управления параметрами технологических процессов, в которых активное участие принимают человек и ЭВМ [2].

Синтез речи относится к группе речевых алгоритмов [3], к которой также можно отнести: распознавание речи, анализ текста на естественном языке, алгоритмы автоматического извлечения информации и знаний из текстов, системы машинного перевода и др. Все эти алгоритмы тесно связаны между собой. Во многих из них применяются сходные технологические подходы. К примеру, задача синтеза речи является обратной к задаче её распознавания. При этом алгоритмы распознавания часто применяются на этапе построения систем синтеза для автоматизированного выделения элементарных звуков речи или сегментных границ.

Группа речевых алгоритмов относится к приоритетным направлениям научных исследований во многих странах мира. Главным образом, это связано со значительными качественными изменениями, которые они привносят и могут привнести в повседневную жизнь человека. Для примера, можно сказать, что только правительство Японии ежегодно тратит порядка 40 млн. долларов на разработки по информатизации японского языка.

Значимость применения речевых алгоритмов при построении человеко-машинных интерфейсов подчёркивает тот факт, что многие современные учёные связывают само появление человека (Homo sapiens) и резкий скачёк в его развитии с появлением у наших предков способности говорить [4]. Можно сказать, что речь – это катализатор развития человечества, позволивший ему за относительно короткое время осуществить резкий скачёк вперёд и получить всё то, что мы имеем сегодня. Главное, что даёт нам речь – это возможность общаться между собой и передавать опыт от одного поколения другому. Общий период эволюции предков человека насчитывает сотни миллионов лет. Возникновение же человека в современном виде (Homo sapiens) и появление у него способности говорить произошло примерно 200 тыс. лет назад. При этом первая письменность, которая возникла на Земле – шумерская. Возникла она примерно от 4 до 5 тысяч лет назад [5]. Из приведённой хронологии видно, что в течение большей части периода своего развития человек пользовался только исключительно устной речью. Отсюда можно сделать следующий вывод: человеческий мозг должен быть более приспособлен к усвоению устно-речевой информации, нежели визуально-текстовой.

В определённой степени речевое восприятие текстовой информации у здорового человека более первично, чем визуальное и стоит выше по иерархии. Главным образом, это связано с тем, что человек изначально учится говорить, а уже только значительно позже приобретает навыки чтения. Существует также проблема неграмотных людей, просто не умеющих читать. Человек приобретает навыки чтения относительно своей уже сформировавшейся и развитой устной речи. В данном случае исключение составляют люди родившиеся глухими. У таких людей процесс обучения чтению происходит путём прямого формирования образных ассоциативных связей с читаемым текстом.

Другая важная особенность: человеческий интеллект устроен таким образом, что способен воспринимать и обрабатывать информацию только в последовательном виде, в виде последовательностей образов или паттернов [6]. Человек не может, моментально увидев лист с печатным текстом, сразу прочитать его и усвоить информацию, содержащуюся в нём. Ему необходимо последовательно слово за словом, предложение за предложением его прочитать. По мере чтения, у него возникают последовательности, состоящие из различных образов. В процессе последовательного чтения у человека также последовательно возникают прямые автоассоциативные образы, связанные с текстом. Общий же смысл текста воспринимается как набор ассоциативных связей более высокого порядка между элементарными образами его составляющими. Для чтения текста человеку также необходимо прикладывать определённые моральные и механические усилия: двигать глазами, фокусировать зрачок, организовывать процесс чтения в рамках установленных порядков и правил.

Иначе происходит процесс восприятия устной речи. Устная речь изначально представлена в последовательной форме в виде последовательностей звуков, слов, предложений. Последовательно также изменяются ударения, интонации и скорость чтения. Восприятие устноречевой информации с помощью слуха происходит у человека практически самопроизвольно на уровне рефлексов. Это подтверждается тем

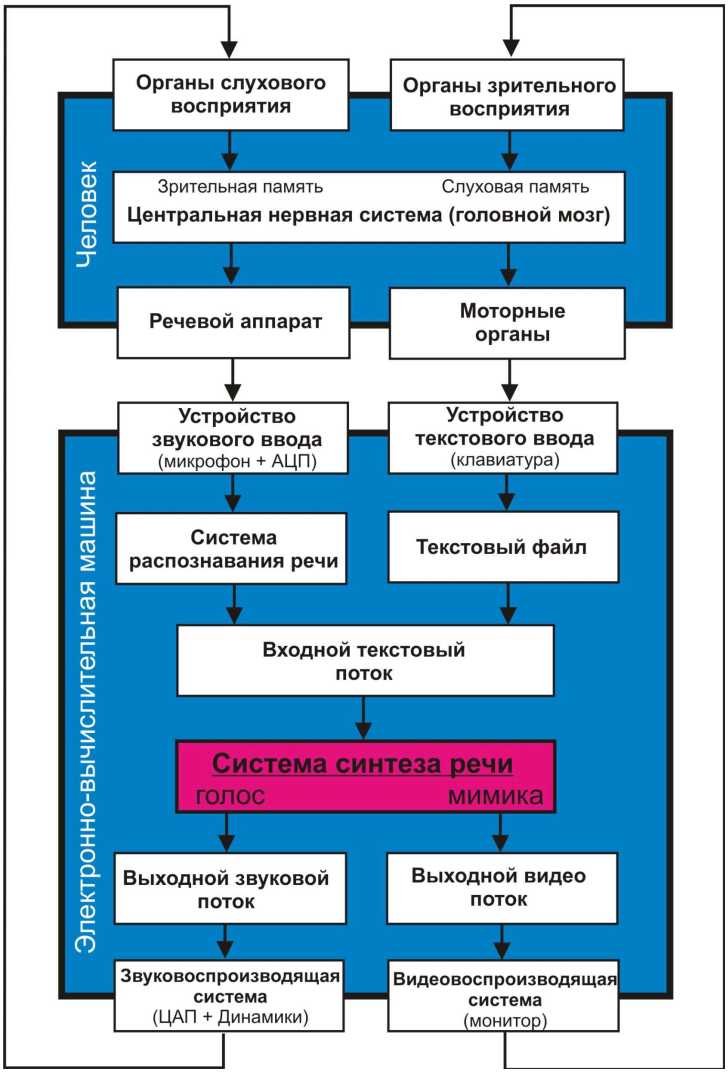

Рис. 1. Схема взаимодействия человека и ЭВМ с применением речевых человеко-машинных интерфейсов.

фактом, что даже если человек не желает воспринимать информационную составляющую речи, восприятие всё равно происходит помимо его воли и единственный способ препятствовать восприятию – это физически устраниться от источника звукового сигнала. В комфортной обстановке при оптимальных параметрах звучания (приемлемый уровень громкости, скорости чтения, отсутствие посторонних шумов) человеку требуются минимальные усилия для восприятия и усвоения информации представленной в устноречевой форме.

Важные различия заключаются также между восприятием устной речи с сопутствующей ей визуальной информацией и восприятием той же речи без сопутствующей визуальной информации. Типичным примером для первого случая является кинофильм, а для второго – радиопередача. В случае просмотра кинофильма человек

мысленно погружается в представляемые ему внешние визуальные и звуковые образы, воспринимая информацию практически такой, какой она ему преподносится. При этом мозговая активность находится на относительно низком уровне. В случае же прослушивания радиопередачи, человеку необходимо «включать» собственное воображение для воссоздания излагаемой картины, ему приходится постоянно извлекать некие образы, хранящиеся в собственной памяти. Другие образы могут возникать самопроизвольно из автоассоциативной памяти. При прослушивании радиопередачи активность головного мозга намного выше, чем при просмотре кинофильма. Данную особенность восприятия человеком информации нужно учитывать при создании обучающих систем с применением ЭВМ и речевых человеко-машинных интерфейсов.

Письменная же речь в общем смысле является «побочным» эффектом устной речи и носит второстепенный характер. Для человека наиболее естественно и эффективно получать и усваивать информацию именно в устной форме. Это связано главным образом с тем, что при устной форме общения между человеком и человеком или между человеком и машиной звуковая информация из слухового нерва попадает непосредственно в ту часть головного мозга человека, которая отвечает за речевые функции. При письменном же общении информация из зрительного нерва сначала попадает в визуальные области головного мозга, обрабатывается там, а уже только потом оттуда попадает в ту часть головного мозга, которая ответственна за речевые функции. Неэффективность второго варианта заключается в наличии большего количества звеньев, через которые должна пройти информация, а также в самом факте задействования менее приспособленной к речевым функциям зрительной памяти. Опыты показывают, что усваивать речевой материал наиболее эффективно задействуя только слуховую память без участия зрительной. В определённой степени зрительная память мешает нам в усвоении речевой информации. Это утверждение главным образом подтверждается общеизвестным фактом, что выучить иностранный язык можно в короткие сроки (от 6 мес. до 1 года) методом «интенсивного погружения в устную языковую среду». Напротив, если изучать язык самостоятельно с применением только письменной литературы (различных словарей и самоучителей), то этот процесс может затянуться на десятилетия, при этом возникает существенный риск неправильного изучения языка и формирования некорректного произношения. Данный подход обосновывает общее применение речевых технологий в информационных системах, а также частное их применение при построении обучающих систем. В данном случае в качестве «обучаемого» может выступать как человек, так и машина. В общем можно сказать, что главным плюсом зрительного восприятия информации является его скорость, а слухового восприятия – более высокая эффективность усвоения и качество самой информации. Человек может быстрым взглядом посмотреть на лист бумаги с печатным текстом, за считанные секунды пробежать глазами между его строк и сразу же мысленно представить себе общий смысл излагаемого в нём текста. Однако при таком быстром чтении возникает риск неправильного или неполного понимания смысла текста, излагаемой в нём информации, а также искажённого её представления.

Особенно эффективно при построении систем с человеко-машинным интерфейсом использовать связку технологий синтеза и распознавания речи. На Рис. 1 видны обратные связи, которые свидетельствуют о возможности осуществления контроля над системой человек-машина при использовании одних только речевых интерфейсов. Например, при условии наличия качественного синтезатора речи ЭВМ может обучать человека иностранному языку и при этом производить контроль правильности произношения иностранных слов с помощью алгоритмов распознавания речи. Авторами данного доклада ведутся разработки подобных систем с целью построения лингвистических тренажеров для обучения английскому и казахскому языкам.

Главная задача, которую приходится решать при реализации вышеописанных систем является задача качественного синтеза речи. Наиболее универсальным решением данной задачи может выступить программная реализация алгоритма синтеза речи в виде универсального модуля с целью последующего его использования в различных системах. Авторами доклада успешно произведена реализация данных алгоритмов для синтеза любых слов на английском языке по их текстовому либо транскрипционному представлению. При этом основной упор делался именно на качество осуществляемого синтеза, а также на широкий диапазон регулирования параметров синтеза. Главным побуждающим мотивом для разработки данной системы послужил произведённый анализ существующих синтезаторов речи и опыт их использования при построении лингвистических тренажеров, который показал их крайне низкое качество и очень узкий диапазон регулируемых параметров (скорость и интонация произношения). Главным образом, данная ситуация объясняется ориентированностью существующих синтезаторов на беглое чтение художественной литературы. При таком подходе многие звуки часто «проглатываются», а слова произносятся нечётко, и нет особой необходимости в широком диапазоне регулирования. Такой подход делает существующие синтезаторы малопригодными для систем обучения иностранным языкам, где на первом плане стоит именно качество синтеза.

В настоящее время для реализации алгоритмов синтеза речи широко применяются электронно-вычислительные машины (ЭВМ). До появления ЭВМ алгоритмы синтеза речи воспроизводились с применением физических моделей речевого тракта человека и продувания через них воздушного потока. Рассматриваемый алгоритм синтеза также не является исключением и реализован с применением программных и аппаратных средств ЭВМ с тем, чтобы задействовать все вытекающие отсюда преимущества:

1. Возможность использования уже имеющегося в наличии широкого спектра существующих программных и аппаратных средств ЭВМ для обработки и анализа звука.
2. Использовать развитые инструментальные средства и новейшие технологии для разработки программного обеспечения (ПО).
3. Производить синтез в режиме реального времени.
4. Передавать результаты синтеза по сетям и каналам связи на значительные расстояния за незначительное время.

5. Встраивать реализацию данного алгоритма в другие системы с целью повышения их функциональности и эффективности.
6. Распространять программную реализацию алгоритма среди миллионов существующих пользователей персональных ЭВМ путём простого копирования.

Так к первому пункту можно отнести множество существующих программ для редактирования оцифрованного звука. Наиболее известные из них: Sound Forge, Cool Edit Pro, Wave Lab. Эти программные средства обладают большим набором функциональных возможностей и позволяют осуществлять качественную обработку и редактирование звука. Их можно использовать для решения следующих задач:

1. Запись исходного звукового материала с микрофона.
2. Обработка и редактирование уже записанного звукового материала.
3. Выделение фрагментов различной длины из исходного звукового материала.
4. Нормализация и регулирование параметров звучания (громкость, скорость, высота и тембр звука).
5. Удаление дефектов звучания и шумов.
6. Наложение различных звуковых эффектов (реверберация, эхо и др.).
7. Сохранение звука в любой из доступных аудио-форматов.
8. Преобразование из одного аудио-формата в другой.
9. Использование средств для частотного и спектрального анализа звуков.

Однако здесь нужно заметить, что при построении систем синтеза речи часто приходится обрабатывать записанную речь, общая продолжительность звучания которой может достигать десятки часов. При этом звукозапись может быть фрагментирована на десятки тысяч аудио-файлов (по предложениям, словам, фонемам и т. п.). И обработка такого огромного массива данных с помощью стандартного ПО является довольно трудоёмким процессом. Данные программы имеет смысл использовать для выделения небольшого и ограниченного количества звуковых фрагментов. Например, при разработке микросегментного синтезатора речи, когда необходимо выделить только ограниченное количество базовых периодических микросегментов речи. Если же стоит задача построения высококачественного аллофонного синтезатора, то для одного только английского языка придётся выделить и проклассифицировать порядка нескольких тысяч звуковых фрагментов. Здесь уже не обойтись без собственной разработки специализированного редактора, который позволяет в автоматическом или в полуавтоматическом режиме производить выделение и классификацию звуковых фрагментов. Программное средство подобного класса было успешно нами разработано и использовано при построении системы дифонного синтеза [1].

Очень интересно решение, которое применяется в современных высококачественных синтезаторах речи на этапе выделения и классификации аллофонной базы: применяется не ручное выделение фрагментов, а задействуются специализированные распознаватели речи, которые автоматически фрагментируют и классифицируют речь диктора по заранее известному тексту и его транскрипции.

Алгоритмы синтеза речи программным методом потенциально могут быть применены практически к любому из языков, на которых говорит современное человечество. Каждый из этих языков в свою очередь имеет собственный словарный запас, лексические, фонетические и граматические особенности. Наиболее оптимальным методом решения задачи синтеза речи может стать разработка специального машинного языка, с помощью которого в универсальной форме будут описываться правила синтеза для каждого из человеческих языков с учётом его специфики.

Нужно сказать, что те вычислительные мощности и аппаратные ресурсы, которыми обладают современные ЭВМ, в достаточной степени позволяют обеспечить качественный синтез речи в режиме реального времени. Основная же сложность возникает на этапе проектирования и программной реализации данных алгоритмов.

Идеальным результатом считается такое преобразование текстовой информации, в результате которого получаемое на выходе звуковое представление речи воспринимается случайно выбранным человеком как естественная речь. В ходе прослушивания искусственно синтезированного звукового фрагмента, у человека не должны самопроизвольно возникать предположения о его неестественности и неразборчивости. Конечно, человек принимающий участие в данном опыте должен быть носителем того же языка, на котором осуществляется синтез. А вся информативная составляющая звукового фрагмента должна быть воспринята и понята человеком в максимально возможной степени. Машина должна максимально приблизиться по качеству к человеку-диктору. Это и есть та самая «идеальная планка», к которой мы стремились при разработчике своей системы синтеза речи.

Однако при оценке качества синтеза, нужно отчётливо понимать, что синтез речи– это результат высшей нервной деятельности человека [6]. И достижение указанных результатов возможно только при использовании алгоритмов аналогичных или подобных по функционалу человеческому мозгу. Пока же, в силу текущих достижений науки и техники мы можем только максимально близко приближаться к указанным результатам по качеству и возможностям синтеза. В системах синтеза речи всегда будут присутствовать вполне определённые ограничения, накладываемые функциональностью используемых алгоритмов и аппаратных мощностей.

Заключение

В работе получены следующие результаты:

1. Разработан программный модуль синтеза произвольных английских слов.
2. Отработаны два различных алгоритма синтеза: дифонный и микросегментный.
3. Реализована система выделения и автоматической классификации дифонов.
4. Разработана уникальная методика обучения иностранным языкам основанная на особенностях звукового восприятия речи человеком.
5. Модуль синтеза речи успешно применён при постороении системы автоматического обучения иностранному языку.

Список литературы

1. *Е.Н. Амиргалиев, Р.Р. Мусабаев.* Информационные технологии искусственного синтеза речи // Вестник КазНПУ. – 2007. – № 4. С. 26-34.
2. *Р.Р. Мусабаев.* Опыт использования технологии синтеза речи при построении системы контроля технологического процесса для нефтеперерабатывающего производства // Проблемы инновационного развития нефтегазовой индустрии: Сб. тр. междунар.науч.-практ. конф. – Алматы: КБТУ, 2008.
3. *Е.Н. Амиргалиев, Р.Р. Мусабаев.* Некоторые направления и задачи обработки лингвистических данных // Вестник КазНТУ. – 2007. – № 6. С. 182-187.
4. *А. Ф. Элфорд.* Загадка возникновения Homo Sapiens. Сборник «Тайное и явное», СПб., 2003 г. С. 182-187.
5. *И. Т. Канева.* Шумерский язык. Центр «Петербургское Востоковедение», 2006 г.
6. *Д. Хокинс, С. Блейксли.* Об интеллекте. Вильямс, 2007 г.

Статья поступила в редакцию 27.04.2008