**Текст презентации НИР**

Цель данной работы – разработка алгоритма для синтеза речи по заранее неизвестному тексту на русском языке.

Для достижения поставленной цели необходимо решить следующие задачи:

* провести анализ существующих методов синтеза речи;
* выбрать наиболее подходящий метод синтеза речи;
* изучить фонетические особенности русского языка;
* собрать базу данных, содержащую необходимый набор исходных элементов для синтеза;
* спроектировать программное обеспечение, реализующее выбранный метод синтеза речи;
* выбрать технические средства для реализации спроектированного программного обеспечения;
* реализовать программное обеспечение;
* провести исследование для оценки качества речи, синтезируемой с помощью разработанного программного обеспечения.

Успешное систем разговорного языка повысит доступность компьютеров и автоматизированных систем для широкого круга пользователей. В настоящее время технология автоматического синтеза находит широкое применение в таких отраслях, как телекоммуникации, мобильные устройства, автомобильная индустрия, компьютеризованные системы, образовательные системы и многих других. Подавляющее большинство систем разговорного языка создаются как универсальные, поддерживающие широкий набор языков. Однако стремление к универсальности может нанести серьезный ущерб качеству синтезируемой речи, потому как каждая языковая система обладает уникальными чертами, которые стираются при попытке создать универсальную систему для всех (или многих) языков.

Многие крупные компании, такие как Mirosoft, Amazon и другие активно занимаются разработкой собственных text-to-speech систем, позволяющих преобразовать в речь неизвестные заранее входные данные (Amazon Polly, IBM Watson Text to Speech, IBM Watson Text to Speech, Google Cloud Text-to-Speech и многие другие). На сегодняшний день можно найти большое количество подобных систем, но все они в основном обладают схожими недостатками:

* не все поддерживают русский язык:
* документация и обучение в основном на английском;
* закрытый исходный код;
* платные;
* нельзя заранее проверить качество синтезированной речи на необходимом тексте.

Проанализировав существующие на рынке решения для преобразования текста в речь, можно сделать вывод, что разработка нового продукта имеет смысл.

Существует несколько алгоритмов генерации речи. Выбор алгоритма зависит, в первую очередь, от задачи, для которой он будет использоваться.

В настоящее время выделяют три группы методов синтеза речи:

* параметрический синтез;
* компилятивный синтез;
* синтез речи по фонетическим правилам.

Параметрический синтез применяется в вокодерных системах. Вокодеры синтезируют речь на основе сигнала, представленного набором непрерывно изменяющихся во времени параметров. Эти параметры и управляют синтезатором речи.

Суть данного метода заключается в компиляции речи из предварительно заготовленных и записанных исходных элементов синтеза – слов, фраз, предложений. Минимальным размером элементов при этом является слово. Таким образом, возможности компилятивного метода ограничены объемом словаря.

Более сложными, но худшими по качеству являются алгоритмы, разделяющие речь на более мелкие части. Метод синтеза речи по фонетическим правилам, основываясь на размере минимальных исходных элементов синтеза, делят на следующие виды:

* микросегментный;
* аллофонный;
* дифонный;
* полуслоговый;
* слоговый;
* синтез из различных единиц разного размера.

В качестве способа генерации речи был выбран синтез по фонетическим правилам, потому как это единственный метод, позволяющий озвучить произвольный заранее неизвестный текст. Для реализации данного подхода необходимо создать фонетико-акустическую базу данных, содержащую большое количество минимальных исходных элементов синтеза.

Выбор исходных единиц синтеза существенно влияет на качество синтезируемой речи, отсюда возникает задача оптимального выбора единиц, с которыми будет работать синтезатор речи.

Необходимо учитывать, что фонетические единицы могут изменяться под действием языковых правил (ассимиляция, диссимиляция, оглушение и изменение гласных звукотипов в зависимости от ударения), а также под действием моторной программы (аккомодация).

При произнесении последовательности звуков возникают переходные участки, так как для произнесения следующего звука необходимо перестроить речевой аппарат. Переходные участки могут быть короткими по длительности, однако несут в себе важную для восприятия речи информацию. Например, информация о твердости или мягкости согласного часто содержится в следующем за ним гласным. Твердость/мягкость согласных в русском языке является смыслоразличительной характеристикой (гроза – грозя). Очевидно, что для качественного синтеза речи необходимо учитывать переходные участки. Для этого необходимо в качестве исходных элементов выбирать более крупные чем отдельные звуки элементы. Такими элементами являются аллофоны.

Кроме аллофонного, существует дифонный синтез речи. Дифоном называют сегмент речи между серединами соседних аллофонов. Однако при дифонном синтезе иногда бывает трудно выделить границу деления на дифоны, что негативно сказывается на качестве синтезируемой речи. Чтобы устранить потерю качества, для каждого дифона необходимо более детально учитывать контекст, что приводит к значительному увеличению размеров исходной базы данных. Использование полуслогов и более крупных элементов в качестве исходных имеет те же недостатки, что и использование дифонов.

Исходя из представленной выше информации, в данной работе в качестве исходных элементов для синтеза будут использоваться аллофоны.

Обычно выделяют и используют два практически обоснованных варианта набора аллофонов: мини- и макси-наборы. В макси-наборах выделяют пять позиционных аллофонов гласных (ударный, частично ударный, первый предударный, не первый предударный, заударный) и двадцать два комбинаторных аллофона гласных, итого для шести гласных русского языка выделяют 2700 аллофонов. Для согласных всего выделяют 1440 аллофонов [6]. В мини-наборе создается 240 аллофонов гласных и 180 аллофонов согласных [6]. Эти оценки являются теоретическими и сильно завышены, поскольку на практике многие аллофоны либо не используются, либо не являются смыслоразличающими, поэтому ими можно пренебречь. В итоге в макси-наборе практически используются 1550 аллофонов гласных и 209 аллофонов согласных, а в мини-наборе – 175 аллофонов гласных и 81 аллофон согласных.

Очевидно, что использование макси-набора представляет собой объемную и сложную задачу, поэтому в данной работе будет использоваться мини-набор. В связи с тем, что для записи базы диктора необходимо производить в студийных условиях, в качестве речевой базы будет использоваться корпус русской устной речи [9], созданный в СПбГУ.