**Тема:** «Исследования разных систем распознавания речи и реализация оптимальной из них для голоса-текстового способа связи»

**Актуальность.** Устная речь является наиболее естественным и простым для человека способом общения. Часто в игровой среде возникает необходимость голоса-текстового вида связи между игроками одной команды. Так же программу могут использовать люди с ограниченными возможностями, которым нужно отправить сообщение. Например люди с плохим зрением или с дефектами рук.

**Цель.** Исследовать системы распознавания речи и с помощью оптимальной реализовать программный продукт для голоса-текстового способа связи.

Общая схема распознавания речи. Сначала на вход нам поступает звуковая волна. Ее мы дроби на маленькие кусочки, фреймы. Длина фрейма - обычно 25 мс, шаг - 10 мс. Они идут с некоторым наложением. После этого из фреймов мы извлекаем наиболее важные признаки. Допустим, нам не важен тембр голоса или пол человека. Мы хотим распознавать язык независимо от этих факторов, так что мы получаем важные признаки. Затем скрытая марковская модель натравливает на все это и выдает на каждом фрейме предсказания, распределение вероятностей по фонем. Скрытая марковская модель пытается угадать, какая именно фонема была сказана на том или ином фрейме. Под конец все это запихивается в граф-декодирования, которое получает распределение вероятностей и учитывает языковую модель. Также учитывается произношение слов и выдаются итоговые гипотезы. В целом, именно так и происходит распознавание речи.

Уровень ошибок в Word (короткий: WER) - это способ измерения производительности .

где

* S - число операций замены слов,
* D – число операций удалений слов из распознанной фразы,
* I - число операций вставки слов
* N - количество слов

где Трасп – время распознавания сигнала, Т – его длительность и измеряется в долях от реального времени.

* Модуль шумоочистки и отделение полезного сигнала.
* Акустическая модель — позволяет оценить распознавание речевого сегмента с точки зрения схожести на звуковом уровне. Для каждого звука изначально строится сложная статистическая модель, которая описывает произнесение этого звука в речи.
* Языковая модель — позволяют определить наиболее вероятные словесные последовательности

На сегодняшний день Sphinx, является самым популярным и работоспособным из открытых движков. Одним из достоинств Sphinx является поддержка множества языков и способность создавать свои собственные модели распознавания речи.

Архитектура Sphinx-4 на верхнем уровне относительно проста. Как показано на рисунке , она включает FrontEnd, клиентскую часть (приложение), декодер и базу знаний.

Блок FrontEnd отвечает за сбор аннотирование и обработку входных данных для чтения с помощью декодера. Кроме того, он извлекает объекты из входных данных для чтения с помощью декодера. Аннотации, предусмотренные во FrontEnd, включают в себя начальный и конечный сегмент данных. Операции, выполняемые на данном этапе, реализуют шумоподавление, автоматическую регулировку усиления, анализ Фурье, спектральную фильтрацию Мэла и др.

База знаний содержит информацию необходимую для декодера. Эта информация включает в себя акустическую модель и модель языка. В свою очередь декодер может послать команду базе знаний, требующую от базы знаний динамически изменять себя на основе результатов поиска. Эти модификации могут заключаться в переключении акустических моделей и/или языка моделей, а также обновлении некоторых параметров, например, дисперсии преобразования для акустических моделей.

База знаний - семантическая модель, описывающая предметную область и позволяющая отвечать на такие вопросы из этой предметной области, ответы на которые в явном виде не присутствуют в базе. База знаний является основным компонентом интеллектуальных и экспертных систем.

Декодер выполняет основную часть работы. Он считывает данные с FrontEnd, сопоставляет их с данными из базы знаний и откликом приложения, а также выполняет поиск в пространстве наиболее вероятных последовательностей слов, которые входят в число претендентов на выбор. Термин «пространство поиска» означает описание наиболее вероятных последовательностей слов, которые динамически обновляются с помощью декодера в процессе декодирования.

В отличие от множества других архитектур распознавателей речи Sphinx-4 позволяет приложению контролировать множество функций речевого движка. Во время декодирования приложение может получать данные от декодера. Эти данные позволяют приложению отслеживать, как происходит процесс декодирования, а также влиять на процесс декодирования до его завершения. Кроме того приложение может обновлять базу знаний в любое время.

Но главным достоинством Sphinx является возможность описания проектируемого распознавателя на уровне формальных моделей, что и послужило основанием для выбора Sphinx-4 в качестве инструментария.

Лингвистическая модель является основой построения приложения распознающего и синтезирующего речь.

Лингвистическая модель состоит из словаря, языковые модели, списка фонем и прочих звуков, обучающего множества. Она служит входными данными для построения акустической модели. В словаре содержится список слов и транскрипции к ним. Транскрипции должны состоять исключительно из фонем, которые присутствуют в списке фонем. Помимо слов есть и другие звуки, не несущие в себе смысловой нагрузки: звуки дыхания, различный шум. Языковая модель – совокупность вероятностей появления слов в речи. В соответствии со всеми вышеперечисленными характеристиками речи записывается аудиобаза и она оформляется в виде обучающих примеров с правильными ответами. Каждая аудиозапись должна иметь своё текстовое представление. Чем больше материала в аудиобазе, тем лучше качество распознавания.

Мел – единица высоты звука, основанная на восприятии этого звука нашими органами слуха.

Кепстр определяет последовательность коэффициентов разложения функции lg [Ф (z)] в степенной ряд.

В соответствии с теорией речеобразования речь представляет собой акустическую волну, которая излучается системой органова затем преобразуется в голосовом тракте. Если предположить, что источники возбуждения и форма голосового тракта относительно независимы, речевой аппарат человека можно представить в виде совокупности генераторов тоновых сигналов и шумов, а также фильтров. Схематично это можно представить так.

PLP (PerceptualLinearPredictive) - метод коэффициентов линейного предсказания. В методе учитываются особенности восприятия различных частот человеком - перед вычислением параметров авторегрессионной модели сигнал проходит определённую предобработку.

Hidden Markov Model (HMM) – статистическая модель, которая может использоваться для решения задачи классификации скрытых параметров на основе наблюдаемых. HMM представляет собой конечный автомат, в котором переходы между состояниями осуществляются с некоторой вероятностью, и задано стартовое состояние, с которого начинается процесс. Через дискретные моменты времени может осуществляться переход в новые состояния. При этом каждому скрытому состоянию с заданной вероятностью соответствует наблюдаемое состояние. Кроме того, текущее состояние автомата зависит только от конечного числа предыдущих, а закон смены состояний не меняется во времени.

Задача 1. Вычисление вероятности последовательности наблюдений.

Задача 2. Нахождение наиболее правдоподобной последовательности скрытых состояний для наблюдаемой последовательности.

Задача 3. Обучение параметров модели по заданной последовательности наблюдений и множеству скрытых состояний.

Алгоритм Баума-Велша.

Также существуют n-граммные модели распознавания речевого потока. Они основаны на предположении, что вероятность употребления очередного слова в предложении зависит только от n-1 слов. Сегодня наиболее популярные биграммные и триграммные модели языка.

**Алгоритм Витерби** — алгоритм поиска наиболее подходящего списка состояний (называемого *путём Витерби*), который в контексте цепей Маркова получает наиболее вероятную последовательность произошедших событий.