УДК 004.5

Булынко С. Ю., Козлова Е. И.

УПРАВЛЕНИЕ ФУНКЦИЯМИ КОМПЬЮТЕРА НА ОСНОВЕ АЛГОРИТМОВ РАСПОЗНАВАНИЯ ГОЛОСОВЫХ СООБЩЕНИЙ

Белорусский государственный университет, г. Минск, Беларусь

Управление машиной с помощью голосовых команд в реальном времени, а также ввод информации посредством человеческой речи может не только упростить жизнь современного человека, но и расширить наши возможности взаимодействия с информационными и киберфизическими системами [1]. В настоящее время технологии распознавания голосовых сообщения используются в таких сферах жизни, как осуществление голосового поиска, голосовое управление навигационными системами, обработка входящих и исходящих звонков, позволяющая клиенту заказывать товары или услуги, отвечать на опросы, получать консультации без участия менеджеров, управление системой “Умный дом”, взаимодействие человека с бытовой техникой и электронными роботами. Цель работы – реализация алгоритмов, способных распознать и перевести в текстовый формат с последующим выполнением команд, введенные человеком посредством голосовых сообщений. Для достижения цели были поставлены задачи проведения анализа исследований в области распознавания голосовых сообщений, выбор методов, разработка и реализация алгоритма распознавания речи.

Получение сообщения и определение формата звукового файла. После получения звукового сигнала при необходимости необходимо его оцифровать, затем перевести для хранения и дальнейшего использования в формат WAV-файла [2, 3], предназначенный для хранения данных оцифрованных аудиосигналов. Он представляет собой две, четко делящиеся, области — заголовок файла и область данных. В заголовке файла хранится информация о размере файла, количестве каналов, частоте дискретизации, количестве бит в сэмпле (глубине звучания). WAV-файл использует стандартную RIFF-структуру, которая группирует содержимое файла из отдельных секций – формат выборок аудиоданных, аудиоданные, и т.п. Каждая секция имеет свой отдельный заголовок и отдельные данные. Заголовок секции указывает на тип секции и количество содержащихся в секции байт. Такой принцип организации позволяет алгоритмам анализировать только необходимые секции, пропуская остальные, которые неизвестны, или которые не требуют обработки. Для WAV-файлов, определено довольно много типов секций, но большинство файлов содержат только две из них – секцию формата ("`fmt`") и секцию данных ("`data`"). Это именно те секции, которые необходимы для описания формата выборок аудиоданных, и для хранения самих аудиоданных. Аудио файл может содержать шум, для устранения которого в работе применена Технология Speech Recognition [1]

Языковая модель. В большинстве подходов принято сначала составлять фонетические транскрипции, а затем отдельной языковой моделью улучшать результат – исправлять грамматические и орфографические ошибки, убирать из расшифровки лишние буквы [4]. Обычно для этих целей применяют технологии нейронных сетей. На вход нейросети подаётся спектрограмма, а на выходе получается матрица распределения вероятностей каждой фонемы по времени. Такую таблицу также называют «emission set». Из выхода «emission set» уже можно получить ответ при помощи алгоритма «greedy decoder» (жадного декодинга), т.е. в каждый момент выбрать наиболее вероятный звук [5].

В ходе проведенного анализа литературных данных установлено, что такой подход не информативен с точки зрения правописания и может дать много ошибок, и чаще применяют метод beam search decoding с использованием перевзвешивания гипотез через языковую модель [5].

Генерация текстового сообщения из голосового. После того как получена таблица emission set, нужно сгенерировать текст. Декодинг — процесс генерации транскрипции по

emission set. Он производится не только по вероятностям, которые нам выдаёт акустическая модель, но и с учётом «мнения» языковой модели. Она может подсказать, насколько вероятно встретить в языке такую комбинацию символов или слов. В данной работе для декодинга использован алгоритм beam search [6]. Его идея заключается в том, что мы не просто выбираем наиболее вероятный звук в отдельный момент, а оцениваем вероятность всей цепочки с учётом уже пройденных слов, храним топ кандидатов на каждом шаге и в итоге выбираем самого вероятного. Причём при выборе кандидатов присваиваем каждому вероятность с учётом ответов и акустической, и языковой моделей.

Исходное представление звукового потока речевого сообщения выглядит как последовательность чисел по времени, а поэтому воспринимается недостаточно информативно. В большинстве случаев используется спектральное представление. Это позволяет разложить звук по волнам разной частоты и узнать, какие волны из исходного звукового потока его формировали и какие характеристики имели. Учитывая логарифмическую зависимость восприятия человеком частот, применяются мел-частотные спектральные коэффициенты. Разные сигналы отличаются по уровню громкости. Чтобы привести аудио к одному виду, нормализуются сигналы и фильтруются высокочастотным фильтром для уменьшения шумов. Pre-emphasis — фильтр для задач распознавания речи. Он усиливает высокие частоты, что повышает устойчивость к шуму и дает больше информации акустической модели.

Исходный сигнал не является стационарным. Он делится на мелкие промежутки (фреймы), перекрывающиеся между собой, которые рассматриваются, как стационарные. К каждому фрейму применяется оконная функция Ханна, чтобы сгладить концы фреймов к нулю. Преобразование Фурье позволяет разложить исходный стационарный сигнал на совокупность гармоник разной частоты и амплитуды. Мы применяем эту операцию к фрейму и получаем его частотное представление. Когда применяем преобразование Фурье ко всем фреймам, формируем спектральное представление. Затем вычисляем мощность спектра. Она равна половине квадрата спектра.

Многочисленные научные исследования показали, что человек распознает низкие частоты лучше, чем высокие, и зависимость его восприятия — логарифмическая. Поэтому к спектру мощности применяется свертка из N-треугольных фильтров с единицей в центре. С увеличением фильтра центр смещается по частоте и логарифмически увеличивается в основании. Это позволяет захватить больше информации в нижних частотах и сжать представление о высоких частотах фрейма. Данные логарифмируются.

В качестве базовой архитектуры используется сверточная нейронная сеть, как наиболее подходящая для этой задачи модель. CHC анализирует пространственные зависимости в изображении через двумерную операцию свертки. Нейросеть анализирует нестационарные сигналы и на основе спектрограммы выявляет важные признаки в частотно-временной области. Чтобы ускорить вычисления и использовать модель на персональном компьютере, создается ограничение при выборе архитектуры. Модель не должна быть слишком глубокой и обладать большим числом обучаемых параметров: это усложняет обучение и увеличивает число операций при прямом проходе.

Алгоритм распознавания голосовых сообщений

1. Запись звукового (речевого) сообщения с устройства ввода (микрофона);
2. Преобразование полученного аудиосообщения в цифровой формат;
3. Выбор наиболее подходящего формата звукового файла для его последующего сохранения и представления в памяти компьютера;
4. Реализация алгоритма beam search decoding с последующей генерацией текста;
5. Обучение нейросети на основе полученного текста с заменой повторяющихся символов одним символом, удалением пропусков и выбором наиболее вероятных слов по словарю;
6. Использование существующих метрик для оценки качества распознавания речи;
7. Передача управляющего воздействия в операционную систему посредством использования API.

На данный момент реализованный алгоритм распознавания речи для управления функциями персонального компьютера способен выполнять следующие функции:

- Распознавание фразы целиком;
- Распознавание отдельных команд;
- Добавление записей в список дел;
- Воспроизведение звуковых файлов;
- Завершение работы компьютера;
- Открытие отдельных приложений;
- Удаление файлов.

Заключение. Основными этапами переноса аналогового звукового сигнала на цифровой носитель являются: аналого-цифровое преобразование и выбор наилучшего формата для аудиофайла, а также извлечение данных для дальнейшей обработки.

В работе рассмотрены и проанализированы алгоритмы распознавания голосовых сообщений: временные динамические алгоритмы, алгоритмы с использованием скрытых Марковских моделей и алгоритмы с использованием искусственных нейронных сетей. В результате анализа работы алгоритмов выбран алгоритм на основе искусственных нейронных сетей, позволяющий каждой новой порции обработанной голосовой информации улучшить качество обработки следующей, уменьшая количество погрешностей.

Реализован алгоритм распознавания голосовых сообщений, позволяющий распознавать как фразы целиком, так и отдельные команды, а также выполнять заданные действия в операционной системе ПК.

Список литературы

1. Распознавание речи. [Электронный ресурс]. – Режим доступа: https://speetech.by/index.php?q=technologies/raspoznovanie. – Дата доступа: 16.12.2022.

2. Формат звуковых файлов WAV. [Электронный ресурс]. – Режим доступа: https://radioprog.ru/post/1025. – Дата доступа: 21.12.2022.

3. Структура WAV файла. [Электронный ресурс]. – Режим доступа: https://audiocoding.ru/articles/2008-05-22-wav-file-structure/. – Дата доступа: 19.12.2022.

4. Как работает распознавание речи. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/vk/blog/579412/. – Дата доступа: 18.12.2022.

5. Understanding greedy search and beam search. [Электронный ресурс]. – Режим доступа: https://medium.com/@jessica_lopez/understanding-greedy-search-and-beam-search-98c1e3cd821d – Дата доступа: 18.12.2022.

6. Проблемы распознавания речи: что еще предстоит решить. [Электронный ресурс]. – Режим доступа: https://apptractor.ru/develop/problemyi-raspoznavaniya-rechi-chto-eshhe-predstoit-reshit.html. – Дата доступа: 18.12.2022.