УДК 004.934

РАЗРАБОТКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ УСТРОЙСТВА СИНХРОННОГО ПЕРЕВОДА ЯЗЫКА ЖЕСТОВ В ЗВУКОВОЙ ФОРМАТ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Семенова В.О.,
Политехнический институт СурГУ
Научный руководитель ***Федоров Д.А.****, к.т.н., доцент,*
Семенова Л.Л.*, ст. преподаватель*
Политехнический институт СурГУ

Эффективность работы системы перевода жестовых слов, состоящей из программного комплекса и устройства считывания и вывода данных, определяется точностью и естественностью работы итогового программного продукта. При этом большая роль в качественной оценке работоспособности отводиться подсистеме «жест-звук», состоящей из двух основных частей: нейронной сети, распознающей жестовое слово и выводящей его значение в текстовой форме, и программы синтеза голоса, прочитывающей выводимый нейронной сетью текст, тем самым имитируя естественную речь человека.

Данная подсистема реализует половину предполагаемого диалогового цикла, ответственного за передачу сообщения от человека с коммуникативными трудностями (рис. 1).

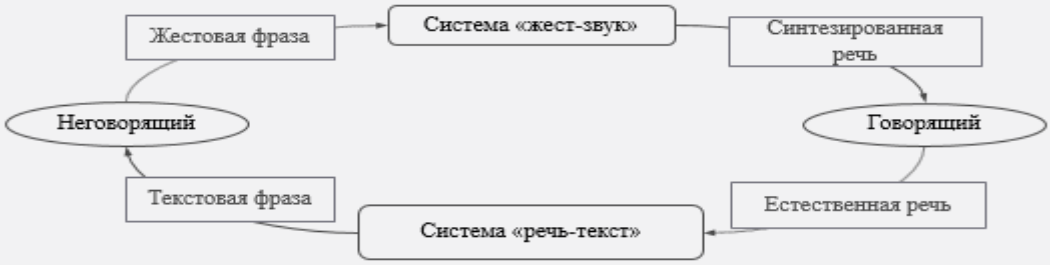

Рис. 1 Схема диалогового цикла с использованием системы

В ходе определения ключевых требований нейронной сети распознавания жестов кислей рук было выявлено, что достаточными требованиями к нейронной сети можно считать следующие: возможность работы с безмаркерной системой и достаточный уровень распознавания жеста кисти руки как при усредненных, так и при отличных от усредненных условиях съемки, включая близкие к граничным (здесь под достаточным уровнем распознавания понимается такой процент считаных жестов, что их количество достаточно для сохранения смысла изначального сообщения) и при низких технических показателях камеры и/или оптических датчиков.

Отдельно ставится требование к корреляции выводов нейронной сети о значении жеста с показаниями гироскопа в связи с большим влиянием на смысловую нагрузку положения кисти руки в пространстве.

По итогу экспериментальных исследований существующих нейронных сетей распознавания жестов кистей рук было решено интегрировать в итоговый программный продукт нейронную сеть с открытым кодом MediaPipe Hands. Данная нейронная сеть соответствует заявленным в проекте требованиям, в последствии возможны изменения исходного кода в случае возникновения необходимости. Распознавание жестов происходит с помощью построения трехмерного скелета ладони, содержащего 21 точку наблюдения, через положение в пространстве которых определяется жест [2]. Использование данной нейронной сети исключает необходимость установки гироскопа и последующей интерпретации его выходных данных благодаря принципу построения скелета, в котором более близкие к камере точки визуально выделяются при распознавании [3].

Для тестирования нейронных сетей на соответствие заявленным требованиям заранее отобранные камеры с различным фокусным расстоянием и качеством съемки были поочередно помещены в условия, имитирующие те, в которых потребитель может использовать устройство: темное время суток/помещение, солнечная погода, туман, загрязнение камеры.

Результаты экспериментального исследования могут быть обобщены в следующих утверждениях:

1) Отобранная нейронная сеть считывает жесты с камеры в реальном времени, при этом отсутствует необходимость хранить в памяти устройства изображения, сделанные в процессе считывания;

2) Отобранная нейронная сеть показывает способность считывать жесты в условиях низкой (0,1 лк) и повышенной освещенности (10000 лк), которые приводят к снижению уровня контрастности между фоном и распознаваемой ладонью;

3) Отобранная нейронная сеть показывает способность считывать жесты в условиях нечеткости границ между фоном и ладонью.

4) Для работы достаточно использования камеры с разрешением не менее 1 МП и углом обзора не менее 100° (фокусное расстояние не менее 20 мм).

Данные выводы свидетельствуют о соответствии выбранной нейронной сети заявленному требованию о достаточном уровне распознавания жестов кисти рук. Полученная в ходе тестирования успешность распознания – 93 %.

В ходе теоретической и экспериментальной исследовательской работы были рассмотрены различные типы подходов к синтезу голоса, а также программные продукты, предназначенные для синтезирования голоса, и были определены основные требования к системе синтеза голоса:

1) Синтезированный голос должен иметь «естественное» звучание, поскольку от этого параметра зависит опыт использования готового устройства потребителем, а также уровень понимания синтезированной речи собеседником;

2) Стоимость использования системы синтеза голоса в случае работы с облачными сервисами не должна иметь сильное влияние на итоговую себестоимость устройства;

3) Необходимо наличие как минимум двух типов голосов – мужского и женского, желательно наличие возможности создания персонализированного голосового модуля для сохранения индивидуальности потребителя при использовании устройства.

По итогу исследовательской работы была выбрана система синтеза голоса Google Text-To-Speech API. Данное решение основано наличие возможности выбора потребителем соотношения цены и «естественности» звучания голоса в зависимости от запроса пользователя, позволяющая, тем самым, регулировать влияние системы на себестоимость устройства через выбор одного из двух профилей, а также возможности интеграции в приложение на смартфон на базе ОС Android при запросе на интеграцию со стороны пользователя [1].

Последовательное использование подсистем, в основе которых лежат отобранные программные продукты, представляет собой программную часть системы распознавания жестовых слов. Иными словами, синтез данных программных компонентов являет собой подсистему «жест-звук».

Список литературы:

1. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search// Cornell University. – URL: https://arxiv.org/abs/2005.11129

2. MediaPipe Hands // MediaPipe – URL: https://google.github.io/mediapipe/solutions/hands

3. MediaPipe Hands: On-device Real-time Hand Tracking // Cornell University – URL: https://arxiv.org/abs/2006.10214