Д.В. Вахнина [1], Д.В. Куликов [1], М.А. Устелемов [1], Н.В. Васильев [1], Е.А. Павлова [1], Ю.А.Егоров [1,2]

[1] Тюменский государственный университет, г. Тюмень

[2] Научно-технический университет «Сириус», г.Сочи

УДК 532.546.2

РАЗРАБОТКА СИСТЕМЫ РАСПОЗНАВАНИЯ РУССКОЯЗЫЧНОГО ТЕКСТА ПО ВИДЕОРЯДУ РЕЧИ

Аннотация. В данной статье представлены разработка датасета фонетической покадровой информации русской речи в видеоряде и исследование моделей классификации изображений губ в момент речи по произносимым фонемам.

Ключевые слова: Классификация, распознавание речи, русская речь, датасет, чтение по губам, фонема, видеоряд.

Введение

Распознавание речи по видеоряду (или чтение по губам) – задача определения произносимого слова, когда за основу берётся только визуальная информация губ говорящего, без сопутствующих аудиоданных.

Целью такого подхода к распознаванию речи является решение проблемы неспособности распознать речь говорящего из-за частичного или полного отсутствия аудиоряда, либо при наличии его дефектов (низкое качество записи, фоновая музыка). Автоматизированное распознавание речи по видеоряду может быть использовано: в приложениях бесшумного набора текста или набора в шумных локациях; в приложениях распознавания речи говорящего, в случае наличия у него затруднения речи (дисфония) или неспособности производить голосовой звук (афония); в видеоредакторах для восстановления поврежденной аудиодорожки.

Задача распознавания речи по видеоряду может быть рассмотрена на двух уровнях:

- Уровень слов: классификация слов из предопределенного набора. Основное число публикаций представлено по решению задачи на данном уровне [1-4]. Проблемой такого подхода является ограниченность датасета определенным набором слов: наибольший из публично доступных датасетов ограничен пятьюстами словами [5].
- Уровень фонем: для каждого кадра видеоряда решение задачи классификации по классам фонем (минимальная звуковая единица языка). По результатам классификации, при таком подходе, потребуется формирование осмысленных слов из получившейся последовательности фонем. Представляет собой универсальных подход, так как для возможности определения слова не требуется его присутствие в обучающей выборке, как это происходит на уровне слов.

Для решения задачи распознавания речи по видеоряду в рамках русского языка потребуется подготовить соответствующий набор данных, так как в связи с отсутствием публикаций по решению данной задачи подходящий датасет в открытом доступе – отсутствует. Для работы был выбран подход к решению задачи на уровне фонем, исходя из его универсальности, что позволит работать с данными в реальных условиях (слова в которых не из обучающей выборки).

Была поставлена цель: разработать систему распознавания русскоязычного текста только по видеоряду записи речи.

Для достижения цели в рамках данной работы были поставлены следующие задачи:

1. Разработать программный инструмент для подготовки датасета фонетической покадровой информации на основе видеоряда;

2. Подготовить датасет с использованием разработанного инструмента;

3. Подготовить словарь соответствий фонемных транскрипций со словами для возможности перевода последовательности фонем в осмысленные слова;

4. Провести исследование моделей классификации изображений губ в момент речи по произносимым фонемам;

5. Проанализировать полученные результаты, определить задачи дальнейшей работы.

1. Разработка инструмента подготовки датасета

1.1. Формирование требований к инструменту подготовки датасета

Требования к интерфейсу разрабатываемого инструмента: настольное приложение Windows.

Требования к организации хранения файлов: входные и выходные файлы хранятся на локальном диске пользователя приложения.

Описание входных данных (два возможных набора исходных данных): файл в mp4-форматес видео- и аудиорядом, а также файл в txt-формате с аннотацией (субтитрами), либо url-адрес видео на видеохостинге YouTube.

Требования к выходным данным:

1. Набор изображений губ в jpg-формате с нормализованным по ключевым точкам положением и соответствующими значениями фонем в названии файла. Каждое изображение соответствует кадру входного видеоряда, кадры, где положение губ не удалось определить – пропускаются;

2. Файл в json-формате с данными по каждому кадру входного видео: номер кадра, фонема, набор координат ключевых точек губ;

3. Файл в mp4-формате с последовательностью кадров губ с нормализованным положением, аннотированных соответствующими

фонемами (текст фонемы наносится на кадр) – необходимо для визуальной проверки работы инструмента сопоставления аудиоряда с субтитрами;

4. Файл в mp4-формате с последовательностью кадров исходного видео ряда, аннотированных соответствующими словами из субтитров – необходимо для проверки субтитров;

5. Файлы, отображающие статистику по полученным данным (графики в png-формате и данные в виде таблиц в txt-формате)

Требования к функционалу:

1. Получение по url-ссылке видео с хостингаYouTube (с указанной стартовой и конечной секундами при необходимости обрезки): исходного видео в mp4-формате, обрезанного видео в mp4-формате, аудиоряда обрезанного видео в wav-формате, субтитров без знаков препинания в txt-формате.

2. Возможность редактирования файла субтитров в окне приложения.

3. Возможность просмотра статистики в окне приложения.

1.2. Описание используемого набора фонем

При разработке инструмента подготовки датасета для использования в качестве модуля сопоставления фонемы с временным интервалом аудиоряда был выбран MAUS [6], так как он является единственным инструментом сопоставления, поддерживающим русский язык на уровне фонем без необходимости дополнительной настройки.В качестве набора фонем MAUS использует нотификацию X-SAMPA (43 фонемы) [7].В таблице 1 приведен список фонем X-SAMPA спримерами использования в словах.

Таблица 1. Список фонем X-SAMPA с примерами использования в словах.

Фонема	Слово	Фонема	Слово	Фонема	Слово	Фонема	Слово
b	Борт	l	Лампа	s_j	Сердце	z	Зуб

b_j	Бюро	l_j	Лес	S_j	Щека	z_j	Зима
d	Дом	m	МаМа	S	Шум	Z	Жена
d_j	ДяДя	m_j	Мяч	t	Точка	a	парА
f	Флаг	n	Нос	t_j	Тётя	e	пЕчь
f_j	Февраль	n_j	НяНя	ts	Царь	i	одИн
g	ноГа	p	ПаПа	tS_	чай	u	мУж
g_j	Герой	p_j	Перо	v	Вор	1	мЫшь
j	дизаЙн	r	Роза	v_j	Верфь	o	абажурнОе
k	Кот	r_j	Рюмка	x	Хор	<:p>	тишина
k_j	Кино	s	Сыр	x_j	Химия		

1.3. Технологии, используемые в разработке инструмента

Языки программирования, используемые в разработке: C#– пользовательский интерфейс десктопного приложения; Python–подготовка и обработка данных, модуль машинного обучения.

Использованы следующие основные технологии и библиотеки:WPF, Scikit-learn, OpenCV, Face-recognition, Docker, MAUSTextAligner.

1.4. Программная реализация инструмента подготовки данных

Для разработки пользовательского интерфейса реализовано 28 функций.

Для модуля обработки данных и машинного обучения реализовано 5 классов и 30 функций.

Интерфейс приложения по подготовке датасета приведён на рисунке 1.

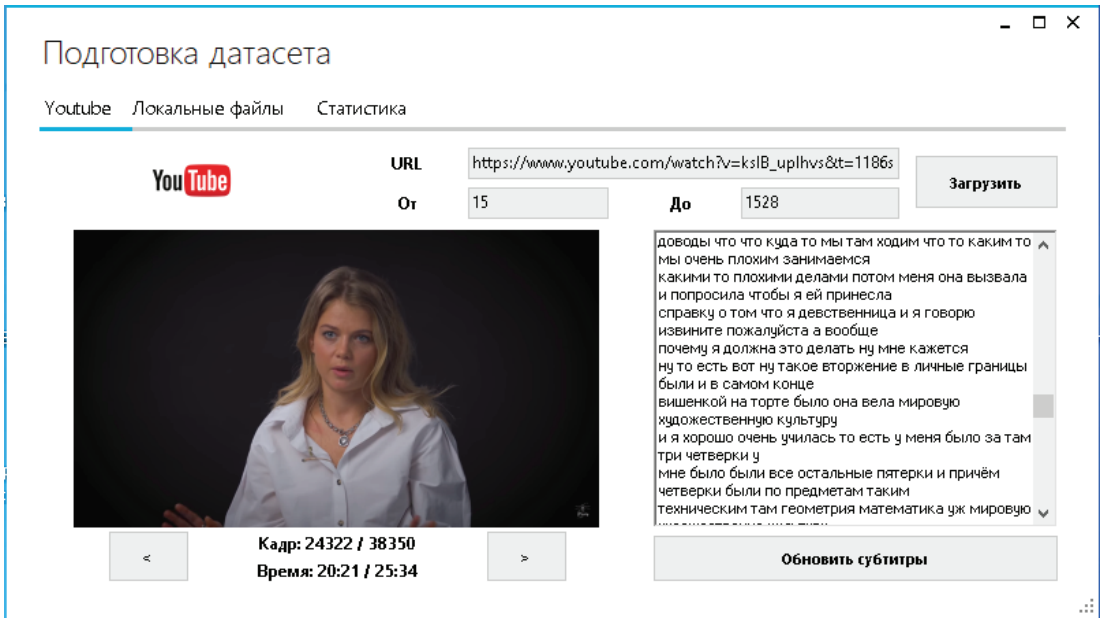

Рис. 1. Интерфейс десктопного приложения по подготовке датасета.

2. Формирование датасета

2.1. Требования к исходным данным

Общие требования: речь одного говорящего; русский язык является родным для говорящего; спонтанная речь; отсутствие густой растительности на лице говорящего; отсутствие речевых и мимических дефектов у говорящего; минимальное число речевых элементов, плохо поддающихся аннотированию (например, «слова-паразиты»).

Требования к видеофайлам: минимальное число монтажных эффектов на кадрах; качество изображения более 480p (854x480 пикселей); съёмка ведётся анфас; число кадров с перекрытием лица говорящего минимально.

Требования к аудиоряду: отсутствие искажений аудиоряда (посторонней речи, фоновой музыки, шумов); высокое качество аудиоряда (частота от 44.1 kHz, битрейт от 128kb/s).

Требования к транскрипции: полное соответствие видеофайлу; допустимо только наличие буквенных символов и пробела.

2.2 Выбор источников исходных данных

Подходящие по требованиям видеоданные были найдены на YouTube-канале «вМесте» [8]. Все видео на канале имеют следующие характеристики:

- интервью одного человека, сидящего анфас;
- закадровая речь отсутствует;
- минимальноечисло монтажных эффектов;
- удаление неинформативных частей речи;
- наличие автоматически сгенерированных субтитров;
- видео в качестве 720p с частотой 25 кадров в секунду.

2.3 Подготовка датасета

На основе описанного источника был подготовлен датасет:

- 5 мужчин и 5 женщин, возрастом от 18 до 35 лет;
- общая длительность видео: 173 минуты 30 секунд (260250 кадров);
- общее количество слов: 28 163;
- число кадров с найденными координатами губ: 260 078.

Распределение количества кадров с найденными координатами губ для разработанного датасета приведено на рисунке 2.

Рис. 2. Распределение количества наборов координат губ по фонемам в подготовленном датасете.

3. Исследование моделей классификации на подготовленных данных

Для проведения исследования моделей классификации в рамках данной работы обучающая и тестовая выборки были сформированы с использованием ключевых точек губ (24 точки, 12 – для нижней губ, 12 – для верхней губы). Были выбраны минимальная и максимальная точки по X и Y координатам (см. рисунок 3).С использованием евклидового расстояния для точек по каждой оси были получены два признака: ширина и высота (соответственно осям).

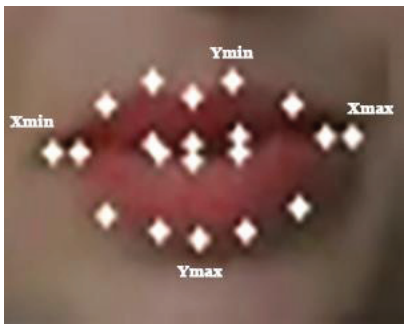

Рис. 3. Ключевые точки губ с граничными значениями по осям

3.1 Проведение классификации

В качестве эталонного результата была рассмотрена работа поклассификации фонем английского языка (41 фонема) с использованием сверточных нейронных сетей [9]. В качестве датасета в работе рассматривался набор из 132 000 кадров (разделение на обучающую и тестовую выборки – 80% и 20% соответственно) видеозаписи одного говорящего. Полученная в работе точность классификации равна 40.3%.

Было проведено исследование для классификации полного набора фонем, набора фонем без фонемы тишины (<p:>), бинарная классификация фонемы тишины и любой другой фонемы (классы: тишина, не тишина), бинарная классификация каждой пары фонем (исключая <:p>).

Результаты точности классификации по различным классификаторам при кросс-валидации со стратификацией приведены в таблице 3.

Таблица 3. Результаты точности классификации различными классификаторами

Модель\Набор данных	Полный список фонем	Без фонемы тишины	Тишина или не тишина	Средняя бинарная классификация
Логистическая регрессия	0.226	**0.257**	0.797	0.748
Линейный дискриминатный анализ	0.226	**0.257**	0.797	0.749
Метод kNN	0.183	0.161	0.768	0.681
Деревья принятия решений	0.151	0.131132	0.752	0.657
Наивный классификатор Байеса	**0.241**	**0.257**	0.789	0.745
Линейный метод опорных векторов	0.225	**0.257**	0.797	0.749
Метод опорных векторов	0.238	**0.257**	0.798	0.75
Многослойный персептрон	**0.241**	**0.257**	**0.799**	**0.752**
Bagging	0.142	0.122	0.751	0.671
Случайный лес	0.145	0.121	0.754	0.673
Экстра-деревья	0.156	0.135	0.759	0.665
Adaboost	0.235	**0.257**	0.798	0.749

3.2. Анализ результатов

Точность классификации по полному набору фонем согласно таблице 3 в сравнении с [9] ниже (лучший показатель – 24.1% в сравнении с 40.3%), что показывает необходимость перехода к работе со сверточными нейронными сетями в качестве модели классификации.

Однако хорошие результаты по бинарной классификации тишины (79.8%) показывают, что даже при текущем подходе имеет место достаточно точное разделение начало и конца слова. При преобразовании последовательности фонем в осмысленные слова низкая точность классификации может быть менее критична, так как вычисление общего расстояния до слова будет способно нивелировать часть ошибок. Для проверки данной гипотезы потребуется разработка инструмента преобразования последовательности фонем в осмысленные слова.

Заключение

В результате работы были выполнены следующие задачи:

1. Разработан программный инструмент для подготовки датасета фонетической покадровой информации на основе видеоряда.

2. С использованием разработанного инструмента подготовлен датасет объёмом 260078 кадров.

3. Подготовлен словарь соответствий фонемных транскрипций со словами для возможности перевода последовательности фонем в осмысленные слова с 1 531 154 слов.

4. Проведено исследование моделей классификации для изображений губ в момент речи по произносимым фонемам. При использовании признаков на основе ключевых точек губ получена точность классификации 24.1% на полном наборе классов.

Проведён анализ полученных результатов, сформулированы задачи для дальнейшей работы:

1. Провести исследование моделей классификации на основе свёрточных нейронных сетей.

2. Разработать инструмент преобразования последовательности фонем в осмысленные слова на основе подготовленного словаря.

3. Разработать конечное приложение для формирования субтитров только по видеоряду.

Благодарности

Статья подготовлена в рамках разработки образовательного кейса для НТУ Сириус при финансовой поддержке РФФИ в рамках научного проекта № 19-37-51028.

СПИСОК ЛИТЕРАТУРЫ

1. Shrivastava N., Saxena A., Kumar Y., Shah R., Mahata D., Stent A. MobiVSR: A Visual Speech Recognition Solution for Mobile Devices [Электронный ресурс] // arXiv. URL: https://arxiv.org/pdf/1905.03968.pdf (дата обращения: 05.06.2020)

2. Saitoh T., Kubokawa M. LiP25w: Word-level Lip Reading Web Application for Smart Device [Электронный ресурс] // AVSP 2019. URL: https://avsp2019.loria.fr/wp-content/uploads/2019/07/AVSP_2019_paper_7.pdf (дата обращения: 04.06.2020)

3. Akbari H., Arora H., Cao L., Mesgarani N. Lip2AudSpec: Speech reconstruction from silent lip movements video [Электронный ресурс] // arXiv. URL: https://arxiv.org/pdf/1710.09798.pdf (дата обращения: 05.06.2020)

4. PetridisS., ShenJ., CetinD. Visual-only recognition of normal, whispered and silent speech [Электронный ресурс] // arXiv. URL: https://arxiv.org/pdf/1802.06399.pdf (дата обращения: 05.06.2020)

5. Chung J., Zisserman A. Lip Reading in the Wild [Электронный ресурс] // Oxford Robotics Institute. URL: http://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16/chung16.pdf (дата обращения: 06.06.2020)

6. WebMAUSBasic [Электронный ресурс] // Bavarian Archive. URL: https://clarin.phonetik.unimuenchen.de/BASWebServices/interface/WebMAUSBasic(дата обращения: 06.06.2020)

7. RussianX-SAMPA [Электронный ресурс] // UCL. URL: https://www.phon.ucl.ac.uk/home/sampa/russian.htm (дата обращения: 06.06.2020)

8. Канал вМесте [Электронный ресурс] //YouTube. URL: https://www.youtube.com/channel/UC9nqcsavbn01KyqZ2d22X8g (дата обращения: 04.06.2020)

9. videoToVoice [Электронный ресурс] //GitHub. URL: https://github.com/carykh/videoToVoice (дата обращения: 04.06.2020)